Scaling Up Learning Models in Public Good Games

Jasmina Arifovic[*]and John Ledyard[†]

December 1, 2003

Abstract

We study three learning rules (reinforcement learning, experience weighted attraction learning, and individual evolutionary learning) and how they perform in three different Groves-Ledyard mechanisms. We are interested in how well these learning rules duplicate human behavior in repeated games with a continuum of strategies. We find that reinforcement learning does not do well, individual evolutionary learning does significantly better, as does experience weighted attraction but only if given a small discretized strategy space,. We identify four main features a learning rule should have in order to stack up against humans in a minimal competency test. Those four are: (1) the use of hypotheticals to create history, (2) the ability to focus only on what is important, (3) the ability to forget history when it is no longer important, and (4) the ability to try new things.

1. Introduction

A significant effort has been spent on building behavioral models of playing games with an eye to explaining laboratory data with human subjects. Both the theory and experiments have been mostly focused on one-shot games with small strategy spaces. Examples include the prisoners' dilemma, matching pennies, chicken, etc. But there is another class of games that is interesting, at least to

[*] Simon Fraser University

[†] California Institute of Technology

those interested in mechanism design. These are repeated games with a one-dimensional continuum as the strategy space. There has been relatively little done in the context of these larger games. The question we address in this paper is whether the learning models that perform well for "small games" scale to "larger games."

We believe the most natural way to scale the study of learning algorithms to environments with large strategy spaces is to begin with a one-dimensional strategy space. The Groves-Ledyard mechanisms in public goods environments creates exactly this type of game. We say "mechanisms" because there is really a class of mechanisms parameterized by a "punishment" parameter. As this parameter changes, the Nash Equilibrium outcome stays the same but the dynamics appear to change. Thus the learning problem, finding the Nash Equilibrium, changes even if the target doesn't. A theory that works for one value of the parameter should, one would hope, work for all. A second way we scale is to look at repeated play. Learning is not only about the game itself but also about one's opponents. But repeated play introduces new intertemporal strategies such as teaching and punishing. We do try to minimize the "repeated game effects" by choosing games for which the Nash Equilibria are Pareto-Optimal.An additional benefit of using this class of games for our study is that there are laboratory data for this class.

For this paper we look at three models of learning: Reinforcement Learning (RL), Experience Weighted Attraction Learning (EWA), and the Individual Evolutionary Learning (IEL)[1]. Two of these are very well known but for completeness we will provide a brief description of each. In RL players choose strategies that have done well in the past with higher probability in the future. Strategies that achieve higher returns when used are "reinforced" and played with a higher probability. In EWA, each strategy has an *attraction* based on the possible payoff it might have earned in the past had it been played. Strategies with higher attractions have higher probabilities of being selected. One of the primary differences between RL and EWA is the latter's use of hypothetical computations to quickly evaluate all strategies. RL only uses actual payoffs and thus can only evaluate strategies actually played. This difference does not matter much when there are two strategies; it is fatal to RL when there are a lot.

[1] We acknowledge up front that neither RL nor EWA were designed to describe behavior in repeated games. They were also not designed to handle games with a continuum of strategies. But the principles of behavior should be independent of the game being played and pushing the boundaries can tell us whether the basics are right or not.

The model of Individual Evolutionary Learning adds a new dimension to RL and EWA by allowing agents to vary their active strategy set in response to experience (as in RL), to hypothetical evaluations (as in EWA), and, occasionally, to pure random events (experiments). What has been "learned" by an agent at any time is summarized not in attraction weights but in a set of active strategies. Strategies that have been or would have been successful will have more copies in the active strategy set. If a strategy has a lot of copies in the active set it will be chosen with a higher probability. The primary difference between IEL and EWA seems to be that IEL discards strategies that aren't potentially profitable and thus does not waste time or lose payoffs re-testing unprofitable options.

To evaluate each of the three learning models we will look at their performance in three Groves-Ledyard mechanisms. We will focus on what we consider to be a minimal competency test. That is, we set a standard of rejection as opposed to one of acceptance. As we simulate the behavior, we compute the average time it takes to converge to the pure strategy Nash Equilibrium. We ask "is it similar to that exhibited by humans?" Our convergence criterion is challenging. We ask that all players' strategies be within 0.1 of the Nash Equilibrium[2]. This is a standard human subjects attain pretty quickly in many experiments. We should expect our models of them to do as well.

One problem we face in doing this test is adapting RL and EWA to the larger strategy space. To do this, we discretize the continuum into a set S. The more elements in S, the more likely the Nash Equilibrium strategies will also be in the set. But the larger is S, the further away we will be from the small set ideal of the learning rules. We will see how this tension plays out below.

We turn now to the study. We begin with a literature review, followed by a review of the Groves-Ledyard mechanisms in this incarnation and of the three learning rules. We then present the results of our study and conclude with some observations and conjectures about the properties of good learning models.

1.1. Some of the past literature

A number of models of individual learning have been developed over the past decade. (For an excellent overview, see Camerer, 2003). Much of the research has been done in the context of one-shot games with small strategy spaces such

[2] Recognizing that getting this close is not necessarily "convergent behavior" since strategies could rebound away, we also look at what percent of the next 100 choices are also within 0.1 of the Nash Equilibrium. If that number is large, we can then think of this as convergence.

as 2 by 2 or 3 by 3 games. The performance of the models has generally been evaluated by using standard econometric methods (maximum likelihood or grid search) to fit the models to experimental data.

Despite of a large number of applications of the models of individual learning to games with a small strategy space, there have been only a few studies examining these models in environments in which there are large strategy spaces or in environments in which there is repeated interaction among the agents. Two studies, Chan and Tang (1998) and Arifovic and Ledyard (2003), have been done in the framework of repeated use of a Groves-Ledyard mechanism for the provision of public goods in an environment with quasi-linear utility functions. This seems to be a particularly good context as a first place to study large strategy spaces and repeated play. It is a large strategy space world, since the mechanism requires the strategy space to be a closed convex subset of the real line and not a finite set. It is a "simple" repeated play world, because the Nash-Equilibrium outcome of the stage game is Pareto-optimal. This significantly reduces repeated game pressures to deviate from Nash Equilibria and allows one to focus on learning as opposed to more complicated coordination, punishment, and teaching strategies.

Chan and Tang (1998) study two major families of learning models (a variant of RL and a Generalized Fictitious Play model). They use the data from the experiments with human subjects to estimate the models' parameter values (using grid search) and evaluate their goodness of fit. They do not perform any out-of-sample tests in order to evaluate the performance of these models. The model that performed the best in their study was RL.

Arifovic and Ledyard (2003) study a model of Individual Evolutionary Learning. Their method for evaluating the performance of the model is different from the above mentioned methods. They examine, in Monte Carlo simulations, the time it takes the model to converge to Nash equilibrium for a wide range of the values of the mechanism's free parameter.

We are interested in ultimately finding learning rules whose parameters can be fixed ex ante across a very large number of games. This is contrary to much of the current literature where parameters are estimated for a limited set of games and, sometimes, even adjusted for different samples of humans. There are exceptions such as the work by Josephson (2001) and Arifovic et al. (2003).

Josephson studies the evolutionary stability of learning rules using numerical analysis. He studies the stability of a class of learning rules that can be represented by EWA (Camerer and Ho, 1999) in four symmetric, 2 player games.

The results of the Monte Carlo simulations show that belief learning is the only learning rule which is evolutionary stable in almost all cases, whereas RL is unstable in almost all cases. In addition, in certain games, the stability of intermediate learning rules hinges critically on the parameters of the model and the relative payoffs. Arifovic et al. (2003) set up a Turing tournament in which the performance of models of individual learning in a number of 2 by 2 and 3 by 3 games is evaluated using machine algorithms designed to differentiate between data generated by human behavior in the controlled laboratory environment and the behavior generated by models of learning.

1.2. Groves-Ledyard Mechanism as a "testbed"[3]

We begin with a very brief review of the Groves-Ledyard (GL) mechanism in quasi-linear environments. We look at both the theory and at some of the experimental evidence.

1.2.1. The environment

We focus on environments in which agents have quasi-linear, quadratic preferences for a public good. An agent's preference for an amount X of the public good is defined for that agent i, $i \in \{1, \ldots, N\}$, as

$$V^i(X) = A^i X - B^i X^2 + \alpha^i.$$

The public good is produced using a constant returns to scale production function with a per unit cost of production, z. Thus the total cost of production is equal to z where z is the total amount of public good that is provided.

1.2.2. The mechanisms

Agents send messages to a *mechanism* (really a central processor or "the government") indicating their demand for the public good. Then, given the vector of messages from the agents, the mechanism computes a level of public good and a tax payment for each agent. Formally we let the set M be the language or message space. Each agent i, $i \in \{1, \ldots, N\}$, selects an element $m^i \in (-\infty, +\infty)$

[3] This section is intended mainly as a reminder to the reader of the structure of the problem. For more details, see Groves and Ledyard (1977), Chen and Plott (1996), or Arifovic and Ledyard (2003).

where m^i is interpreted to be the agent's message to the government. The total amount of public good produced is given by:

$$X(m) = \sum_{i=1}^{N} m^i.$$

The message m^i sent by agent i, $i \in \{1, \ldots, N\}$, can be thought of as representing that agent's requested addition to the total amount of public good (given the proposed additions of other agents). Agents are free to misrepresent their requests for the public good and, if this were a voluntary mechanism, we would expect them to do so. However, the tax and allocation rules of the mechanism are specifically designed so that in Nash equilibrium it is in each agent's individual self-interest to reveal her true incremental demand for the public good. The GL mechanisms, parameterized by γ, use the following tax scheme:

$$T^i(m) = (X(m)/N)z + (\gamma/2)\left[\frac{N-1}{N}\left(m^i - \mu^i\right)^2 - \sigma^{i2}\right]$$

where T^i is the amount of tax paid by agent i, γ is an arbitrary free parameter greater than 0, $\mu^i = \frac{\sum_{h \neq i} m^h}{N-1}$ is the mean value of messages of all the other agents, and $\sigma^{i2} = \frac{\sum_{h \neq i}(m^h - \mu^i)^2}{N-2}$ is the squared deviation from this mean. We call

$$g(m) = \left(X(m), T^1(m), ..., T^N(m)|\gamma\right)$$

the GL outcome function. The payoff of agent i, if the messages are m, is

$$U^i(m) = V^i(X(m)) - T^i(m)$$

1.2.3. One-shot play

This is an incentive compatible mechanism with a balanced budget on and off the equilibrium path. It is well known that, in this environment with quasi-linear preferences, the Nash equilibrium public good outcome of the one-shot game will be the unique, Pareto optimal level of public good.[4] So, in particular, in quasi-linear environments, the Nash-equilibrium outcome level of the public good is independent of γ .

[4] In more general environments, there can be multiple Pareto-optimal allocations. A Nash equilibrium of the Groves-Ledyard mechanism will select one of these.

1.2.4. Repeated play

In a repeated play version of the public good allocation problem, it is assumed that the public good lasts only for 1 period. Further, payoffs are additive over time without discounting. So at each iteration t, an amount X_t of the public good and taxes, T_t^i are chosen. An agent's payoff from the sequence $(X_1, T_1, ..., X_{t'}, T_{t'})$ is

$$\begin{aligned} U^{*i} &= \sum_{t=1}^{t'} U^i(m_t) \\ &= \sum_{t=1}^{t'} V^i(X_t) - T_t^i. \end{aligned}$$

It can be shown, at least for agents following Cournot best response strategies, that γ is important for the dynamic performance of the mechanism. Chen and Tang (1998) derive a sufficient condition for the convergence of the mechanism in repeated play in which agents play best responses given the messages of the other agents.[5] If agents use best responses in a sequence of repeated stage GL mechanisms, messages will converge to Nash equilibrium if agents' strategies are strategic complements; i.e., if in the stage game

$$\partial^2 U^i / \partial m^i \partial m^j \geq 0.$$

This is true for quadratic preferences iff $\gamma \geq 2NB^i$ for all i. Thus, the strategic complementarity condition is satisfied for a sufficiently high value of γ. For the set of the parameter values in Chen and Tang 1998 which we use in this paper, $\gamma \geq 2NB^i$ for all i holds for values of γ greater than 80.

1.3. The experimental data

The major purpose of learning models is to explain the data from controlled experiments in which human subjects are confronted with various mechanisms. There are several data sources for the GL mechanism. One set is described in Chen and Plott (1996). They conducted 7 experimental sessions each with $\gamma = 1$ and $\gamma = 100$. The second set is described in Arifovic and Ledyard (2003). It was generated by us at the California Institute of Technology in April and May

[5] See also, Muench and Walker 1983 and Page and Tassier 2003 for further analysis of related dynamics.

2002. We conducted 4 experimental sessions with $\gamma = 50$ and 3 experimental sessions with $\gamma = 150$. [6]

We are interested in several questions. Does convergence to the predicted Nash Equilibrium messages occur? If so, how fast? There are many possible measures of performance one might use to answer these questions. In this paper, a convergence criterion is defined in terms of how close all agents' messages are to the equilibrium messages. The convergence occurs in the period when the difference between the equilibrium value and the value of the selected message of each agent is less than or equal, in absolute terms, to 0.1; i.e., when $|m_i^a - m_i^e| \leq 0.1$ for all i. The period when the convergence criterion is fulfilled is called *the time of the first passage through equilibrium*, $T_c^{\gamma,r}$ for run r and given γ. The average time of the first passage through equilibrium for R runs, $\bar{T}_c^{\gamma}$ is given by:

$$\bar{T}_c^{\gamma} = \frac{\sum_{r=1}^{R} T_c^{\gamma,r}}{R}. \tag{1.1}$$

We denote the standard deviation from this value, across the R runs, by $\sigma_{T_c^{\gamma}}$.

How *stable* is a Nash equilibrium after the first passage? We have created a measure called the *index of equilibrium stability* E_s^{γ}. It measures the frequencies with which equilibrium values of messages are represented in the entire sets of agents' response during periods after the first passage through equilibrium. Our measure of the stability of equilibrium is

$$E_c^{\gamma} = \frac{\sum_{t=T_c^{\gamma}+1}^{T_{max}} \sum_{i=1}^{N} E_i(t)}{N(T_{max} - T_c^{\gamma})} \tag{1.2}$$

where $E_i(t)$ is an index variable such that $E_i(t) = 1$ if $m_i(t)$, the message that subject i sent at experimental period t is equal to m_i^e and $E_i(t) = 0$ if $m_i(t) \neq m_i^e$, and T_{max} is a total number of periods in a given experiment session. The average over a total number of experimental sessions conducted for each γ and multiplied by 100 is given by E_s^{γ}. We average over the remaining number of periods of a particular session once the first passage through equilibrium is achieved.

Data from the experiments can be found in Table 1.

Table 1

[6] Our experimental design is very similar to Chen's and Tang's. However, we introduced two modifications. First, in Chen and Tang, subjects could make only integer number choices. We let the subjects make real number choices, with a two decimal points restriction. Second, we added a calculator to the windows interface that allowed the subjects to calculate their potential payoffs varying the size of μ^I and σ^{I2}.

Convergence Times and Stability of Equilibria

γ	*observations*	$\bar{T}^c_\gamma$ ($\sigma_{T^c_\gamma}$)	E^γ_s ($\sigma_{E^\gamma_s}$)
50	3	5.75 (4.42)	98.00 (1.00)
100	7	18.86 (12.034)	99.00 (1.00)
150	4	20.00 (17.20)	92.00 (4.00)

We will refer back to these data later. For now perhaps the most important observations are that convergence occurs relatively rapidly on average, that the fastest convergence occurs for $\gamma = 50$, and that beyond 50, a value for which the strategic complementarity condition does not hold, convergence times increase in γ. We want to see whether learning models can duplicate this type of behavior.

2. The Learning Models

We look at three learning models that all have some features in common. At each point in time t, they each involve a *finite* set of strategies, S^i_t, and a probability, π^i_t, of playing each strategy from that set. That is, each model employs a mixed strategy as a basic component. The probabilities, π^i_t, are based on a cumulative, but depreciated, reckoning of the payoff that strategy would have received at each past chance. The key computation is the (expected) payoff that the alternative $a^i_{j,t}$ received when it was used or would have received if it had been actually used, taking the behavior of other agents as given.[7] The behavior of the others, their messages, can be summarized in two statistics: $\mu^i = \frac{\sum_{h \neq i} mh}{N-1}$, the mean value of messages of all the other agents, and $\sigma^{i2} = \frac{\sum_{h \neq i} (mh - \mu^i)^2}{N-2}$, the squared deviation from this mean. Given, $s^i_t = (\mu^i_t, \sigma^{i2}_t)$ an agent can compute

$$X(a^i_j, \mu^i_t) = a^i_j + (N-1)\mu^i_t$$

for each alternative $a^i_j \in A^i_t$. Then they can also compute

$$\begin{aligned} U^i(a^i_j | s^i_t) &= V_i(X(a^i_j, \mu^i_t)) - T_i(a^i_j, \mu^i_t, \sigma^{i2}_t) \\ &= V_i(X(a^i_j, \mu^i_t)) - X(a^i_j, \mu^i_t)z - (\gamma/2)\left(\left(\frac{N-1}{N}\right)\left(a^i_j - \mu^i_t\right)^2 - \sigma^{i2}_t\right) \end{aligned}$$

$U^i(a^i_j | s^i_t)$ is i's foregone utility for a^i_j. given that others played s^i_t.

[7] This is an entirely retrospective and myopic view of the situation an agent faces. We ignore for now intertemporal strategies (e.g., grim triggers, tit-for-tat, etc.), and other complexities introduced by repeated play. We intend to address these issues in our future research.

2.1. Reinforcement Learning

Reinforcement Learning models trace their origins to the psychological learning theories of Thurstone (1930), Bush and Mosteller (1955), and Luce (1959). They have been adopted by many economists including Cross (1983), Roth and Erev (1995, 1998), and Bender, Mookherjee, and Ray (1998). Their models assume that players choose strategies probabilisticly. Players choose those strategies that have done well in the past with higher probability in the future. Thus strategies that achieve higher returns are "reinforced" and will be played with a higher probability.

We follow Chen and Tang's implementation of an adjusted RL algorithm. First the continuous strategy space is discretized. A finite set of active strategies S is determined. This set will remain fixed throughout a run of the mechanism. In Chen and Tang's experiments, subjects could choose an integer, $5x \in \{-20, \ldots, 30\}$. Thus each agent had 51 active stage-game strategies. They divided each choice number by 5 and rounded it up to the nearest integer. This way, the number of strategies was reduced to 11. In our simulations below we will use both 11 and 51 as sizes for the sets S.

Each active strategy in S has a *propensity of choice* that depends on the past payoffs earned by a given strategy and that determines the strategy's probability of being selected. Propensities of choice of different strategies are updated based on the payoff earned in a round of the game when a particular strategy was used and is otherwise left at its previous round level. Strategies are selected based on their propensities. Those with higher propensities have higher probabilities of being selected.

There is one main variable that is updated after each round of experience. $R_{ij}(t)$ is the attractiveness of strategy j to i at t. $R_{ij}(t)$ begins with a prior value, $R_{ij}(0)$. These initial values can represent both prior game experience and/or player predictions. For each agent, i, and each strategy j in S, let $I_{ij}(s_t^i)$ denote the indicator function where $I_{ij}(s_t^i)$ is equal to 1 if alternative (strategy) j is chosen at round t in the experiment, and 0 otherwise. The propensity of choice, $R_{ij}(t)$, begins at $R_{ij}(0)$ and is updated according to

$$R_{ij}(t) = [q \cdot R_{ij}(t-1)] + \left[I_{ij}(s_t^i) \cdot U^i(a_j^i | s_t^i)\right] \tag{2.1}$$

where $q \in [0, 1]$ is a time/memory discount factor.[8]

[8] Notice that there is no need to calculate foregone utilities in this model.

Agent i selects strategy j at t+1 with probability:

$$\pi_{ij}(t+1) = \frac{e^{\lambda R_{ij}(t)}}{\sum_{k=1}^{|S|} e^{\lambda R_{ik}(t)}} \tag{2.2}$$

for every i and j where $|S|$ is a total number of strategies in S. We will call λ a focus parameter since it determines the extent to which the agent focuses on choices with higher values of $R_{ij}(t)$. If $\lambda = 0$ then $\pi_{ij}(t+1) = 1/|S|$ for all ij and there is no focus. As $\lambda \to \infty$, $\pi_{ij}(t+1) \to 1$ for the j such that[9] $R_{ij}(t) > R_{ij'}(t)$ for all j'. At $\lambda = \infty$, there is total focus on the strategy with the largest value of R_k. The free parameters of this model are $S, q, R_{ij}(0)$, and λ.

2.2. Experience Weighted Attraction Learning

EWA generalizes the RL model of the previous section by allowing agents to weigh hypothetical payoffs as well as actual. We follow Camerer and Ho (1999) to describe the version of EWA that we implement for this paper. First, the continuous strategy space is discretized. A finite set of active strategies S is determined. In our simulations below we will use both 11 and 51 as sizes for the sets S.

Each active strategy in S has an attraction, called the propensity of choice in the RL model, that depends on the past payoffs earned by a given strategy and that determines the strategy's probability of being selected. The attractions of different strategies are updated based on the possible payoff a strategy might have earned had it been played. Strategies are selected based on their attraction. Those with higher attractions have higher probabilities of being selected. There are two main variables that are updated after each round of experience: $N(t)$, the number of "observation-equivalents" of past experience; and $A_{ik}(t)$, player i's attraction to strategy k after period t has taken place. $N(t)$ and $A_{ik}(t)$ begin with some prior values, $N(0)$ and $A_{ik}(0)$. These initial values can represent both prior game experience and/or player predictions.

The experience weight starts at $N(0)$ and is updated according to

$$N(t) = \rho N(t-1) + 1$$

[9] If such a j does not exist then the probability is spread evenly across all the j such that $R_{ij}(t) \geq R_{ij'}(t)$.

for $t \geq 1$ where ρ is a depreciation rate or retrospective discount factor. This means that

$$N(t) = \left(\frac{1}{1-\rho}\right) + \left(N(0) - \frac{1}{1-\rho}\right)\rho^t.$$

Note that for $\rho = 0$, $N(t) = 1$ for all t. And for $\rho \to 1$, $N(t) \to N(0) + t$. Let[10]

$$L_{ik}(t) = \left[\frac{\delta + (1-\delta)I_{ik}(s_t^i)}{N(t)}\right] = \left[\frac{\delta + (1-\delta)I_{ik}(s_t^i)}{\left(\frac{1}{1-\rho}\right) + \left(N(0) - \frac{1}{1-\rho}\right)\rho^t}\right].$$

The attraction of a strategy k for player i at time $t+1$ is updated from $A_{ik}(t)$ given s_t^i, according to

$$A_{ik}(t+1) = [\phi \cdot A_{ik}(t)] + \left[L_{ik}(t, s_t^i) \cdot U^i(a_k^i | s_t^i)\right] \tag{2.3}$$

The parameter δ determines the extent to which hypothetical evaluations will be used in computing attractions. If $\delta = 0$ then no hypotheticals are used, just as in RL. If $\delta = 1$ then hypothetical evaluations are fully weighted. The factor ϕ is a discount factor or decay rate, which depreciates the previous attraction. ϕ is similar to q in the RL model.

Agent i selects strategy j at t+1, in exactly the same way as with RL, with probability:

$$\pi_{ij}(t+1) = \frac{e^{\lambda A_{ij}(t)}}{\sum_{k=1}^{|S|} e^{\lambda A_{ik}(t)}} \tag{2.4}$$

The free parameters of this model are: $|S|, N(0), A^j(0), \rho, \phi, \delta$, and λ.

We end this section by pointing out the implications for a couple of specific parameter values. If $\phi = q, \delta = 0, \rho = 0$, and $N(0) = 0$ then this is just the RL model. If $\lambda \to \infty$, and $q = 0$, then this is a best-reply model and all the probability is put on the strategy that maximizes utility in response to s_t^i.

2.3. Individual Evolutionary Learning

IEL adds a new dimension to standard learning models by allowing agents to vary their active strategy set in response to experience, to hypothetical evaluations, and, occasionally, to pure random events (experimentation). What has been "learned" by i at time t is summarized not in attraction weights but in the

[10] The indicator function $I_{ik}(s_t^i)$ is defined in the previous section.

set of active strategies. We follow Arifovic and Ledyard (2003) to describe the version of IEL we use in this paper.

At the beginning of round t, each agent $i \in [1, \ldots, N]$ has a collection S_t^i of active strategies. S_t^i consists of J alternatives[11] where $a_{j,t}^i \in M$, for $j \in \{1, \ldots, J\}$. In each round each agent computes a new S_{t+1}^i. In order for S_t^i to accumulate and retain "good strategies", there must be a way to try out almost any strategy in the original large set, $(-\infty, +\infty)$. Experimentation is the way new strategies are added by IEL. But there must also be a way to purge strategies from S_t^i that are not likely to provide a good payoff. Replication is the way old, low-payoff strategies are purged by IEL.

Experimentation works as follows. For each $j = 1, ..., J$, with probability ρ select one message at random from M and replace $a_{j,t}^i$ with that message.[12] This, apparently random, experimentation introduces new alternatives that otherwise might not ever have a chance to be tried. But, the result of this experimentation is not as random as it looks. While it is true that an alternative is selected at random from M, we will see that the alternative selected must also have a reasonably high foregone utility relative to the last period or future periods to have any chance of ever being used. [13]

After changing S_t^i with experimentation, we further modify S_{t+1}^i using replication to reinforce messages that would have been good choices in previous rounds using hypothetical foregone utility computations. We allow potentially better paying alternatives to replace those that might pay less. For $j = 1, \ldots, J$, we let $a_{j,t+1}^i$ be chosen as follows. Pick two members of S_t^i randomly (with

[11] J is a free parameter of the behavioral model. It can be loosely thought of as a measure of the processing and/or memory capacity of the agent.

[12] For the simulations in this paper we used a rate of experimentation $\rho = 0.033$. For the random selection of the replacement, we used a normal density with the mean equal to the value of the alternative, $a_{j,t}^i$ that was to be replaced by a 'new' idea. The standard deviation was set to 1.

[13] There are at least two possible interpretations of our experimentation process. One is that it is a *trembling hand mistake* and the other is that it is *purposeful experimentation* intended to improve an agent's payoff. We feel the latter interpretation is most appropriate because a choice generated through experimentation is implemented only if it demonstrates a potential for bringing a higher payoff. Thus, we call this method *directed experimentation* since only those newly generated alternatives that appear promising are actually tried out.

uniform probability) with replacement. Let these be $a^i_{k,t}$ and $a^i_{l,t}$. Then[14] let

$$a^i_{j,t+1} = \left\{ \begin{array}{c} a^i_{k,t} \\ a^i_{l,t} \end{array} \right\} if \left\{ \begin{array}{c} U^i(a^i_{k,t}|s^i_t) \geq U^i(a^i_{l,t}|s^i_t) \\ U^i(a^i_{k,t}|s^i_t) < U^i(a^i_{l,t}|s^i_t) \end{array} \right\}.$$

Replication at $t+1$ favors alternatives with a lot of replicates at t and alternatives that would have paid well at t if they had been used. So it is a process with a form of averaging over past periods - if the actual messages of others have provided a favorable situation for an alternative $a^i_{j,t}$ on average then that alternative will tend to accumulate replicates in A^i_t, (it is fondly remembered), and thus will be more likely to be actually used in future moves. If the responses of the others are fairly stable, then over time, the sets A^i_t will become more homogeneous as most alternatives become replicates of the best performing alternative.

Given A^i_{t+1}, the selection probabilities are updated by letting

$$\pi^i_k(t+1) = \frac{U^i(a^i_{k,t+1}|s^i_t) + \varepsilon^i_{t+1}}{\sum_{j=1}^J (U^i(a^i_{j,t+1}|s^i_t) + \varepsilon^i_{t+1})} \tag{2.5}$$

for all $i \in \{1, \ldots, N\}$ and $k \in \{1, ..., J\}$ and where[15]

$$\varepsilon^i_{t+1} = \underset{a^\epsilon \in A^i_{t+1}}{min} \; \{0, U^i(a^\epsilon|s^i_t)\}.$$

Over time, copies of a (hypothetically) successful strategy accrue in the active strategy set S^i_t, and this increases the probability that strategy will be selected even if its utility is only slightly larger than some other. The reader may be curious why we used the expression in (2.5) instead of the more common one in (2.4). It turns out that based on the extensive simulation in Arifovic and Ledyard (2003) it really does not matter much which we choose. For the simulation results we report below, we used the set of parameter values that, at 95% confidence interval, resulted in the mean value and the variance of the time of first passage through equilibrium closest to the experiments with human subjects. The expression (2.5) was part of that accepted model. So we carried it over to here.

The free parameters for this model [16] are: J, A^i_o, and ρ.

[14] We could have used the $R_{ij}(t)$ from reinforcement learning or the $A_{ij}(t)$ from EWA in place of $U(a^i_{jj,t}|s_t)$.

[15] Using ε to make sure probabilities are positive has another helpful effect. when ε becomes more negative, more probability is assigned, ceteris paribus, to k with large $U(a^\epsilon|s_t)$ and less is assigned to k with very negative $U(a^\epsilon|s_t)$.

[16] One might also consider other things to be free parameters such as the method of initially

3. Simulation Results

We are, ultimately, interested in discovering which, if any, of the three learning rules provides a good model of human subject behavior in experiments. Our initial test for each is something we consider a minimal competency test. We ask whether the learning rules exhibit, in simulations, the same convergence behavior as human subjects.[17]

3.1. The Basics: convergence and stability

For each of the three learning rules and each of the three values of $\gamma = 50, 100, 150$, we ran 100 simulations. Each simulation was allowed to run for $10,000$ periods. We also carried out a number of simulations with variations in the parameters to check that our results were robust to our choices. For our robustness simulations we also used 100 runs with 10,000 periods. These are described below in the appendix. Rather than report all data, we use several summary statistics to characterize the behavior of the simulations. The primary measures provided are **convergence times and equilibrium stability.** These were defined earlier in the section on experimental data. The average time to convergence, and its variance, were defined above in (1.1) and our measure of equilibrium stability was defined in (1.2). We also present figures that demonstrate the evolution of these measures over 200 periods.

We will first provide a separate analysis of each rule and then conclude with some comparisons and explanations.

3.1.1. Reinforcement Learning

Chen and Tang found the set of the RL parameters that gave the best fit with the experimental data using a grid search, were $\lambda = 0.006$, and $q = 0.8$. We conducted our baseline RL simulations using those parameter values and a strategy space size $|S| = 11$. For these values, RL strategies did converge to the Nash Equilibrium. However, the amount of time it takes for this to happen is two

seeding A_0^i, the form of the probability used in selecting messages during replication, and the way we do random selection from A_t^i at the end. In Arifovic and Ledyard (2003) we report on a extensive list of robustness tests we did in which these were all varied in significant ways. The performance of the algorithm did not seem to depend too seriously on the particular way.

[17] We should be clear on this. Passing this test is a necessary condition for saying the rule behaves like humans do. Passing the test is certainly not sufficient for such a conclusion..

orders of magnitude larger than the time it takes in the experiments. In Table 2 we report the average convergence times and standard deviations for each of the simulations. The column *observations* reports the number of simulations, out of total of 100, that converged for each value of γ within 10,000 periods. Averages $\bar{T}_\gamma^c$ are then computed using the data only from those simulations that converged. We extended the simulation time for another 100 periods in those simulations where we observed convergence and collected data on individual agents' selection of messages. We use that data to compute the index of equilibrium stability, also reported in Table 2.

Table 2
Convergence Times and Stability of Equilibria
RL with $|S| = 11$, $q = 0.8$

γ	observations	$\bar{T}_\gamma^c$ $(\sigma_{T_\gamma^c})$	E_s^γ $(\sigma_{E_s^\gamma})$
50	93/100	3281.55(2365.14)	96.49(3.71)
100	98/100	1686.78(1712.44)	97.04(2.74)
150	100	1665.43(1829.24)	97.73(1.23)

In figure 1, we present the behavior of actual messages chosen by agents over the first 200 periods, for $|S| = 11$, $\lambda = 0.06$ and $q = 0.8$. Agents' messages fluctuate between different values within the first 200 periods.[18] Before period 200, agent 2 settles on $m_t^2 = 0$, agent 3 on $m_t^3 = 2$ and agent 5 on $m_t^5 = 1$. The simulation reaches equilibrium values of messages for all of the 5 agents in period 2776.

[insert fig 1 here]

It seems pretty clear that, although RL does eventually converge to equilibrium behavior, it does not look anything like the behavior of laboratory subjects. The time RL takes is two orders of magnitude greater than humans.

[18] For $\gamma = 100$, the values of equilibrium messages are: $m_e^1 = 0.96$, $m_e^2 = 1.04$, $m_e^3 = 0.98$, $m_e^4 = 1.02$, and $m =_e^5= 1$. Given that, according to our convergence criterion, equilibrium is achieved if the difference between actual and equilibrium messages is less or equal to one, the values of messages of all 5 agents that are equal to 1 would satisfy the convergence criterion.

Robustness To make sure that this very long average time to convergence is not just the result of a strange choice of parameters, we did a number of simulations[19] across systematic variations in q and λ. The basic finding did not change much. An increase in q to 0.9 leaves only 1/100 simulations converging. Lowering q below 0.8 slows things down relative to $q = 0.8$. We also checked values of $\lambda = 0.06$ and 0.6 and found in both cases that no convergence occurred. For a wide range of parameters, RL converges (or learns) a lot slower than experimental subjects do in this experimental environment.

Another robustness test involved increasing the size of the message space. It was not obvious to us, initially, how such an increase would affect average times to convergence. The increase in size could slow things down by creating many more alternatives that need to be evaluated. On the other hand, by including more strategies "near" the equilibrium strategy[20], it might be possible for the learning rule to focus more effort in that neighborhood. As it turns out the former effect dominates the latter for RL. When the strategy space was increased to $|S| = 51$, none of the simulations that we conducted using the RL algorithm converged to the Nash equilibrium values within $10,000$ periods. Figure 2 shows the messages selected by agents over the first 200 periods for $\gamma = 100$, $|S| = 51$ and $q = 0.8$ in a typical one of our simulations. In this simulation, after initial adjustment, agent 1 settles to selecting the value of 1.2. Messages selected by agent 2 fluctuate widely, in the range from -4 to 5. After a lot of fluctuation, agent 3 settles on the value of 1.6, agent 4 to 1.8 and agent 5 to 0.8. However, an examination of the entire simulation ($10,000$ periods) indicates that there is no convergence to equilibrium or non-equilibrium values. For all 5 agents, there are recurrent intervals of wide fluctuations in the values of selected messages, followed by temporary settlement to specific values.

[insert fig 2 here]

To check the robustness of these results for $|S| = 51$, we did a number of simulations across systematic variations in q and λ. The basic findings did not change. We used 3 different values of the parameter q, 0.9, 0.7, and 0.5 with $\lambda = 0.006$. None of the simulations resulted in convergence to equilibrium within

[19] These robustness simulations are only summarized here. We provide considerably more detail in the Appendix.

[20] In fact increasing the size of the strategy space might also cause the equilibrium strategy to become one of the possible choices. For example, for $\gamma \leq 25$, an equally partitioned strategy space of size 11 will not include the equilibrium strategy for those mechanisms.

10,000 periods. We also used values of $\lambda = 0.06$ and $\lambda = 0.6$. None of these simulations resulted in the algorithm's convergence.

The conclusion is that larger strategy spaces lead to even slower average times to convergence for RL. $|S| = 51$ is too big for this learning rule to process in a timely fashion.

3.1.2. Experience Weighted Attraction Learning

Camerer and Ho (1999) report on various estimation results that give the best fit to experimental data for different games. They report values $\rho = \{.961, .946, .935, .926\}$, $\phi = \{1.040, 1.005, .986, .991\}$, $\delta = \{0, .73, .413, .547\}$, $\lambda = \{.508, .182, .646, .218\}$, $N(0) = \{19.63,\ 18.391, 15.276, 9.937\}$. For our baseline simulations with EWA, we used the parameter values of $\lambda = 0.35$, $N_0 = 10$, $\delta = 0.9$, $\rho = 0.95$, and[21] $\phi = 0.9$. We also used $|S| = 11$. For these values, we observed relatively fast convergence for $\gamma = 50, 100, 150$. We report the average time of first passage through equilibrium and standard deviations in table 3.

Table 3
Convergence Times and Stability of Equilibria
EWA with $|S| = 11$, $\delta = 0.9, \phi = 0.9, \lambda = 0.35$

γ	simulations	$\bar{T}^c_\gamma$ $(\sigma_{T^c_\gamma})$
50	100	12.1(3.64)
100	100	12.9(6.98)
150	100	20.26(16.53)

In figure 3, we report the actual messages selected by agents over the first 200 periods. Agents' messages reach the equilibrium value relatively fast, and by period 100, all of the agents are playing only the equilibrium strategies.

[insert fig 3]

EWA does converge to equilibrium strategies and, at least for $\gamma = 100$ and 150, does so in average times that look a lot like those generated by humans. For $\gamma = 50$, however, EWA appears to be slower than does the human data.

[21] We initially chose $\phi = .99$ is accord with the numbers above. But then as we did our robustness studies, we found that, at least in the context of repeated GL games, EWA performed much closer to humans when $\phi = .9$. We decided to go with the parameter values that performed better. The appendix has some of the comparative data in Table 15 across alternative values of ϕ.

Robustness To see whether there was anything special about these results, we did a number of simulations[22] across systematic variations in λ, δ, ϕ, and N_0. Lower values of δ resulted in a substantial decrease in the number of simulations that converged to equilibrium. A ten-fold decrease in the value of λ (0.035) did not have an impact on the convergence times of simulations with $\delta = 0.9$. However, it did result in a tremendous increase in the number of simulations that converged for lower value of δ. Moving ϕ up from 0.9 or down from 0.7 slowed things down. Dropping ϕ to .5 or less led to no simulations converging. Finally to test the effect of a change in the value of N_0, we tried a ten times lower value of $N_0 = 1$. This did not seem to affect the algorithm's behavior.

As with RL we wanted to see what the effect of increasing the size of the message space would be. Unfortunately, when the strategy space is increased to $|S| = 51$, only a small number of EWA simulations converged to the equilibrium messages in 10,000 periods. We report the average convergence times and our measure of stability of equilibrium in table 4.

Table 4
Convergence Times and Stability of Equilibria
EWA with $|S| = 51$

γ	simulations	$\bar{T}^c_\gamma$ ($\sigma_{T^c_\gamma}$)	E^γ_s ($\sigma_{E^\gamma_s}$)
50	0		
100	4/100	31.50 (2.89)	99.95(0.1)
150	1/100	31	100 (00)

In figure 4, we report the actual messages selected by agents over the first 200 periods. Agent 1 is the only one that starts playing the equilibrium strategy, while the other 4 converge very quickly (period 40) to non-equilibrium values, $m_1 = 1$, $m_2 = 0.6$, $m_3 = 0.8$, $m_4 = 0.8$, and $m_5 = 0.6$. All the agents remain with these selected messages until the end of simulation (period 10,000).

[insert fig 4]

The conclusion for EWA is the same as that for RL - larger strategy spaces lead to slower convergence. $|S| = 51$ is too big for this learning rule to process in a timely fashion.

[22] These are summarized here and provided in more detail in the Appendix

3.1.3. Individual Evolutionary Learning

Arifovic and Ledyard (2003) found that the parameters that gave the best fit, at a 95% confidence interval, resulted in the mean value and the variance of the time of first passage through equilibrium closest to the experiments with human subjects. These values are $J = 100$ and $\rho_{ex} = 0.33$ with experimentation process drawn from the normal distribution. In table 5, we report the values of the average time of first passage through equilibrium (and its standard deviation) and the stability of equilibrium (and its standard deviation) for the values of $\gamma = 50, 100$, and 150.

Table 5 Convergence Times and Stability of Equilibria [23]

IEL with $J = 100$, $\rho_{ex} = 0.33$

γ	simulations	$\bar{T}^c_\gamma$ $(\sigma_{T^c_\gamma})$	E^γ_s $(\sigma_{E^\gamma_s})$
50	10000	13.48 (5.76)	95.31 (1.21)
100	10000	19.65 (10.52)	95.00 (1.74)
150	10000	38.89 (22.81)	94.72 (2.09)

Figure 5 reports the behavior of the messages selected by players for the first 100 periods in a simulation with $\gamma = 100$. Relatively fast convergence to the equilibrium values occurs. Occasional out-of-equilibrium values are observed as a result of continuous effects of experimentation. However, the stability of equilibria, once reached is always above 95 %.

[insert figure 5 here]

IEL does converge to equilibrium strategies and, at least for $\gamma = 50$ and 100, does so in average times that look a lot like those generated by humans. For $\gamma = 150$, however, IEL appears to be slower than does the human data.

Robustness We did a significant number of alternative simulations with different parameter values, as well as changed the details of the model's updating scheme. These are reported in detail in Arifovic and Ledyard (2003). The model always yielded the same pattern of convergence to Nash equilibrium where the time to convergence depended on the values of the parameter γ. It always yielded a U-shaped curve (in γ) of times to convergence, where different parameter sets resulted in higher or lower values of the average time of first passage through

[23] These results are taken from Arifovic and Ledyard (2003).

equilibrium. The range of values of γ between 40 and 50 always resulted in the minimum time to convergence.[24]

3.2. Summary to here

At this point we can make several observations about the conformance of the three learning rules with human behavior in the GL mechanisms in the quadratic environment we have used. Numerical comparisons are provided in Table 6. RL appears to have failed our minimum competency test - average times to convergence are not even close to those of humans, no matter what the size of the strategy space. IEL seems to pass the test. EWA provides us with a bit of a dilemma. It appears to pass the test when $|S| = 11$ but not when $|S| = 51$.[25] We need to look deeper to see what is really happening for the larger discretized strategy space.

Table 6 Summary of Average Convergence Times

γ	Experiment	RL (11)	RL (51)	EWA (11)	EWA (51)	IEL
50	5.75	3282	*DNC*	12.1(4)	*DNC*	13.48(6)
100	18.86	1687	*DNC*	12.9(7)	*DNC*	19.65(10)
150	20.00	1665	*DNC*	20.16(17)	*DNC*	38.89(22)

DNC: Did Not Converge in most of the simulations

3.3. A Deeper Look: other measures of convergence

We have been focused on the convergence of realized strategy choices to equilibrium pure strategies. This measure, however, ignores what is going on with the rest of the active strategies. If convergence does not occur, or does not occur as rapidly as in the human experiments, it might be because of the various randomizations that the learning models employ. That is, the mixed strategies could be converging to "reasonable" strategies, even if the realizations were off. We would not want to reject a model of behavior for which this was true

[24] Experiments with human subjects also showed fastest convergence for $\gamma = 50$.

[25] This suggests that one might always be able to find a value for $|S|$ ex post that explains the data. We do not feel that this is enough. It is important to be able to choose S ex ante if we want to predict behavior in the lab.

3.3.1. Difference in mixed strategy distributions

To check out how the learning rules are doing in the mixed strategy space, we consider the entire cumulative distribution of the mixed strategies being played. In particular, we compute a distance between the current mixed strategy, $\pi^i_{t,}$ and the equilibrium strategy, π^i_e. The cumulative distribution for any finite mixed strategy[26] π on the real line $(-\infty, +\infty)$ is given by $F(x|\pi) = \sum_{y \leq x, F(y|\pi)>0} \pi(y)$. We use the following pseudo-metric:[27]

$$
\begin{array}{rcl}
d(\pi^i_t, \pi^i_e) & = & \sum\limits_{x \in Z} |F(x|\pi^i_t) - F(x|\pi^i_e)|/|Z| \\
 & & \text{where} \\
Z & = & \{x|F(x|\pi^i_t) > 0\} \cup \{x|F(x|\pi^i_e) > 0\} \\
 & & \text{and} \\
|Z| & = & \#\text{ elements in Z.}
\end{array}
$$

Since we observed convergence to equilibrium for EWA with $|S| = 11$ and for IEL, we should also observe the difference in cumulative strategy distributions converge to zero. The interesting question is what happens for RL, for both strategy space sizes, and for EWA with a strategy space size of 51. We look at those now.

Reinforcement Learning In figure 6 we present our distance measure for RL with a strategy space size of $|S| = 11$. Until period 200 this measure stays above zero for all of the agents except agent 5. For this agent, the measure drops to 0 fairly quickly, after period 60, and remains there until the end of the simulation. Other agents' measures remain at positive values until the convergence, at period 2776 occurs. Between periods 2776 and $10,000$, this measure remains at the values close to 0 most of the time. In Figure 7 we show the same measure for RL with $|S| = 51$. Until period 200 this measure stays above zero for all of the

[26] A finite mixed strategy π takes on positive values $\pi(x) > 0$ at only a finite number of $x \in (-\infty, +\infty)$.

[27] If the cumulative distributions were piece-wise continuous the distance measure would be the Stieljes intergral $\int_X |F(x|\pi^i_t) - F(x|\pi^i_e)|dx,$. the area between the two distributions. In the data below, the distances between the x with positive probability are uniform so |Z| is the appropriate normalization constant. This is called a pseudo-metric since $d(\pi, \pi') = 0$ does not imply $\pi = \pi'$ as is required of a metric.

agents. Only the value for agent 1 takes a really small positive value. Similarly as with the messages that agents selected over the same period of time, the pattern that this measure displays are recurrent intervals of fluctuations followed by temporary settlement to the values in the range between 0.5 and 0.04.

[insert fig 6]
[insert fig 7]

This confirms the view that convergence is just really slow for RL.

Experience Weighted Attraction Learning In figure 8, we show the difference in cumulative distributions for EWA with $|S| = 11$. This distance takes a zero value around the time when agents strategies converge to equilibrium values, and when their probabilities of playing equilibrium strategies also go to 1. At that point, differences in cumulative distributions go to zero. This as it should be since convergence to equilibrium implies convergence to zero for the difference in cumulative distributions. In figure 9, we show the same measure for $|S| = 51$. This figure is interesting as it demonstrates that the difference in cumulative distributions converge to positive values close to zero. (For agent 1 who converges to playing an equilibrium strategy it does go to 0.) These values remain unchanged until the end of the simulation.

[insert fig 8]
[insert fig 9]

So even though EWA is not converging to the equilibrium, it is getting it almost right.

3.3.2. Efficiency

It should be realized that, even if convergence is not occurring in the pure strategy space or in the mixed strategy space, it is possible that the losses sustained by the agents might not be very large - that convergence in utility is occurring even if convergence in mixed strategies is not. Because we are in a quasi-linear environment and because we are using the GL mechanism, the maximum total payoff to the agents occurs at the Nash equilibrium. A standard measure of how far off the maximum joint payoff the agents are is called **efficiency** and it is computed in the following way:

$$\frac{\sum_i U^i(a^i_{k,t}|s^i_t)}{\sum_i U^i(a^i_e|s^i_e)} \tag{3.1}$$

where $U^i(a^i_{k,t}|s^i_t)$ is the utility of agent i at time t that resulted from her actual choice, a^i_t and the actual choices of the others, s^i_t. $U^i(a^i_e|s^i_e)$ is the utility of agent i if all agents, including i, send equilibrium messages. For our environments, $\sum_i U^i(a^i_e|s^i_e) > 0$ so efficiency may be negative but it is always less than or equal to 1.

If strategies are converging to Nash Equilibrium, then by design efficiency is converging to 1. So again the interesting cases are for RL for both strategy space sizes and for EWA for a strategy space size of 51. We look at those now.

Reinforcement Learning In Figure 10 we display the value of efficiency for a typical simulation of RL with $|S| = 11$.

[insert fig 10]

This looks pretty good in that there are a lot of 1.0's although there are occasional drops down to 0.7 or lower. Efficiency is lower on average and fluctuates more widely. Over the span of 200 moves, the average level of efficiency is 0.78 with a standard deviation of 0.45. Of course, in the first 50 moves (which are important for comparison with the experimental data) the average efficiency is only 0.52.

However, when we look at the same measure for $|S| = 51$, we get a much different story. Efficiency is lower on average and fluctuates more widely. Even after 200 moves, there is no apparent settling down in the fluctuations. Over the span of 200 moves, the average efficiency is 0.61 and the standard deviation is 0.56. Thus, as the size of the strategy space increases, RL begins to look less and less like the experimental data. These fluctuations are the result of fluctuations in individual messages, as demonstrated in figure 2. Again, the same pattern (or lack of it?) characterizes simulations of RL when $|S| = 51$.

[insert fig 11]

Experience Weighted Attraction Learning In figure 12, we report the measure of efficiency for a typical simulation for EWA with $|S| = 11$. As the simulation converges to equilibrium values relatively fast, the efficiency reaches the value of 1 after period 60 and remains at that level until period 200 as expected. We report the behavior of the efficiency computed for EWA and $|S| = 51$ in figure 13. After initial adjustment, convergence to the value of 0.91 occurs. This is a simulation during which 1 out of 5 agents converge to equilibrium messages. In other simulations[28] with δ around .9, more agents converge yielding efficiencies ranging from .97 to 1. This mirrors the observation that EWA is getting close to but not always converging to equilibrium. The large strategy space is causing trouble, although not nearly as much as it does for RL.

[insert fig 12]
[insert fig 13]

3.4. Summary

Our goal was to study what happens to some learning rules when we apply them to mechanisms with a continuum of strategies - a scaling up of the size of the problem they are asked to deal with. We have looked at three: Reinforcement Learning, Experience Weighted Attraction, and Individual Evolutionary Learning. We ask how they compare to human data on the same problem.

We found that RL is not able to handle the larger problem. RL is simply much slower in evaluating strategies than humans are.

We found that EWA did pretty well. When we used a discretized strategy space of 11 options, EWA did indeed converge to equilibrium in a reasonable amount of time, especially for $\gamma = 150$ and 100. But when we increased the discretized space to 51 options, EWA rarely converged. It too is apparently challenged by a larger strategy space. However, when we looked deeper we found that although strict convergence in pure strategies rarely occurred, convergence in mixed strategies and efficiency did seem to be taking place.[29]

When we look across the three rules (focusing on the $|S| = 11$ data) and the human data (see Table 6), we see that, for $\gamma = 50$, EWA and IEL are close in performance and closer to the human data than RL. For larger $\gamma = 100$ EWA seems to be closer to the human data that IEL and for $\gamma = 150$ IEL seems

[28] See the appendix for details.

[29] We do not have a good comparison to the human data for the rates of convergence of efficiency. That will be done in future work.

to perform closer to the human data than does EWA. Both however easily lie within two standard deviations of the human data. Clearly, more work needs to be done before making any more of these observations.

4. Conclusions and Final Thoughts

In three different repeated Groves-Ledyard mechanisms, humans converge pretty fast to Nash Equilibrium strategies (taking from 5-20 iterations). Once they get there, they stay (92-98% of the next plays are also near equilibrium). The three learning rules we studied (RL, EWA, and IEL) match that performance with varying degrees of success. Reinforcement Learning does not match it at all. Experience Weighted Attraction Learning does well for a small strategy space but not for a large space. Individual Evolutionary Learning seems to match pretty well. But all the rules seem to be slower than humans. What explains the variance in performance? Can we find any hints in the data for improvement?

We believe that the key to matching the performance of humans in games with large strategy sets is the quick (endogenous) discovery of "good" strategies and the quick discarding of "bad" strategies. One key to the quick discovery of good strategies from a large set is **the extensive use of hypotheticals.** If one only evaluates one strategy per round, then it takes a while to learn which strategies might work. In 20 rounds, the maximum time it takes humans to converge on average, with 51 strategies and a uniform starting probability of choice, you will only have a probability of .4 of trying the Nash Equilibrium just once. With 5 people, the probability of all getting near is extremely low. It does not improve much even if the strategy space size is 11. Using hypotheticals, one evaluates every strategy in every round. With 11 strategies, the first order effect is to reduce the rounds needed to converge by an order of magnitude. But there must be a second order effect (since EWA is two orders of magnitude faster than RL) and we conjecture this arises because all 5 are moving faster towards the equilibrium and reinforcing each other. Not only is the use of hypotheticals supported by the relative performance of EWA and RL, it is also supported by the robustness tests we did on EWA. There the parameter δ determines the weight given to hypotheticals. We saw that higher δ led to faster convergence - convergence that looked more like the human data[30].

[30] It is interesting to note that Josephsen (2001) found the most evolutionarily stable rule had $\delta = 1$, its maximum value.

It is our belief, based on the simulations in this paper, that the reason EWA and IEL do as well as they do, and better than RL, relative to human data is their use of hypotheticals.

A second key to the quick discovery of good strategies is the ability to **focus on good ones when you find them**. Each learning rule uses a probabilistic choice process. For RL and EWA, a parameter λ determines how important the attraction or utility of a strategy is in this choice. The higher λ, the more important the attraction is. So one might expect that very high λ would be good. But there is a down-side to that choice. High λ inhibits the ability of the rule to actually try a strategy with a low attraction and, thereby, verify its performance is as predicted by the hypothetical. If it does not do such checking, it will converge but perhaps not to the Nash Equilibrium. So, in the RL or EWA models, low λ allow full analysis of the strategy space but create slow convergence by forcing consideration and use of provably bad strategies. High λ speed convergence but inhibit full analysis of the strategy space. The larger the strategy space is, the worse these effects are.

In the IEL model, λ appears to play no role. In the robustness tests done in Arifovic and Ledyard (2003) we found that varying λ led to virtually no change in the average times to convergence. The reason is straight-forward. By maintaining an active set of strategies, as opposed to all strategies, IEL forces a probability zero of choice on those not included. That is, if a strategy is discarded it is never tried. This is the equivalent of a very large λ, especially when the active strategy set becomes populated with many copies of a very few good strategies. But initially, the active strategy set can have many different strategies and, even when there are few strategies, experimentation offers the possibility to try all strategies. This is the equivalent of a very small λ. So early on, and randomly later on, IEL acts as if it is using probabilistic choice with a small λ, thus sampling lots of possible strategies. Later on, and pretty soon too, IEL acts as if it is using probabilistic choice with a very large λ, thus focusing on provenly good strategies.

It is our belief, based on the simulations in this paper, that the reason IEL does better than EWA, especially when a strategy space of size 51 is used for EWA, is its ability to act as if it is adjusting the focus parameter, λ, over time.

Another key to convergence is **the ability to forget history** once it has been absorbed. This requires a delicate balance in the choice of q for RL and the choice of ϕ for EWA. Too high and convergence is slowed down by blind allegiance to the past. Too low and convergence is slowed down because the past is forgotten before it is fully used. Further, this memory parameter is applied uniformly to all strategies. As we saw, $q = 0.8$ and $\phi = 0.9$ seem to work best respectively for RL and EWA in these games. IEU deals with this issue through its process of tournament replication. By randomly selecting challengers to pair off against each other, those strategies which historically have been successful, actually or hypothetically, have a high chance to stay around, while those which have historically been unsuccessful will disappear. Thus only the important parts of history are preserved.

It is our belief, based on the simulations in this paper, that the reason IEL does better than EWA (and RL) is its ability to selectively remember and discard history.

Finally, playing a role in all of this is the size of the strategy space. We have forced RL and EWA to pick the size and content of active strategies ex ante and exogenously. We found that the performance of these learning rules relative to the human data deteriorated as the size of the active strategy space grew. This is at odds with the desire to have a finer strategy grid to better approximate Nash Equilibrium strategies. IEU chooses the size ex ante and exogenously, but the content of **the active strategy set is chosen endogenously over time**[31]. By experimentation (mutation), all strategies in the continuum have a chance to be admitted and retained in this set.

It is our belief, based on the simulations in this paper, that the reason IEL does better than EWA (and RL) is its ability to endogenously adjust the number and content of the active strategy set.

So it seems to us that good learning rules, rules that behave as human subjects do, will have at least four components: (1) the use of hypotheticals to create history, (2) the ability to focus only on what is important, (3) the ability to forget history when it is no longer important, and (4) the ability to try new things. RL has almost none of these characteristics, EWA and IEl have some of them. Further research is necessary, however, before accepting any particular rule as the best. For example, it is likely that something else - such as expectations formation and teaching - will be needed to explain repeated cooperation in games in which the Nash Equilibrium is not Pareto-Optimal.

[31] This also means that its "size" - the number of distinct strategies - is chosen over time.

References

Arifovic, J. and J. Ledyard (2003), “Computer Testbeds and Mechanism Design: Application to the Class of Groves-Ledyard Mechanisms for Provision of Public Goods”, mimeo.

Arifovic, J., R., McKelvey, and S. Pevnitskaya (2003), ”An Initial Implementation of the Turing Tournament to Learning in Two Person Games”, mimeo.

Bender, J., D. Mookherjee, and D. Ray (1998). ”Reinforcement Learning in Repeated Games”, mimeo.

Bush, R. and F. Mosteller (1955). Stochastic Models For Learning. New York: Wiley.

Camerer, C. and T. Ho (1999) ”Experience Weighted Attraction Learning in Normal Form Games”, Econometrica 67: 827-873.

Ho, T., C. Camerer, and J. Chong (2002). ”Functional EWA - A one Parameter Theory of Learning in Games”, mimeo.

Camerer, C.F., (2003) Behavioral Game Theory: Experiments in Strategic Interaction, Princeton University Press.

Chen, Y. and C. Plott (1996) “The Groves-Ledyard Mechanism: An Experimental Study of Insitutional Design.” Journal of Public Economics 59: 335-364.

Chen, Y. and F. Tang (1998) “Learning and Incentive Compatible Mechanisms for Public Goods Provision: An Experimental Study”, Journal of Political Economy 106: 633-662.

Cross, J.G. (1983). A Theory of Adaptive Behavior, Cambridge Universtiy Pess.

Erev, I. and A. Roth (1998) ”Predicting How People Play Games: Reinforcement Le arning in Experimental Games with Unique, Mixed Strategy Equilibria”, American Economic Review 80, 848 - 881.

Groves, T. and J. Leydard (1977) “Optimal Allocation of Public Goods: A Solution to the ‘Free Rider’ Problem”, Econometrica, 45: 783-809.

Josephson, J. (2001) ”A Numerical Analysis of the Evolutionary Stability of Learning Rules”, working paper, available at

http://swopec.hss.se/hastef/abs/hastef0474.htm

Ledyard, J. (1995) ”Public Goods: A Survey of Experimental Research”, in J. Kagel and A. Roth (eds.) *Handbook of Experimental Economics*, pp. 111-194. Princeton: Princeton University Press.

McAllister, P. (1991). ”Adaptive Approaches to Stochastic Programming”. Annals of Operations Research 30: 45-6

Roth, A. and I. Erev (1995) ”Learning in Extensive Games: Experimental Data and Simple Dynamic Model in the Intermediate Term" *Games and Economic Behavior* 8: 164-212.

Page, S.E. and T. Tassier (2003) "The Existence and Stability of Equilibria in the Groves Ledyard Mechanism", manuscript.

Thurstone, L.L. (1930) "The Learning Function", Jouranl of General Psychology 3, 469-493.

5. Appendix

Here we report more of the details of the simulations we ran, in order to check the robustness of our results to changes in parameters of the models.

5.1. Robustness of Reinforcement Learning simulations

We first report on the behavior of RL with $|S| = 11$ when the value of q is varied. For $q = 0.9$, 1 out of 100 simulations converged to equilibrium (and that one admittedly fast, in 40 periods). However, no simulation converged to equilibrium for $\gamma = 50$, and again only one did for $\gamma = 150$, in 873 periods. Decreasing the value of q by 0.1, to 0.7 resulted in worse performance (relative to simulations with $q = 0.8$) in terms of the number of simulations that converged to equilibrium. These results are reported in table 7.

Table 7 RL with $|S|$ = 11, $q = 0.7$

γ	observations	$\bar{T}^c_\gamma$ $(\sigma_{T^c_\gamma})$	E^γ_s $(\sigma_{E^\gamma_s})$
50	0		
100	85	1686.78(1712.44)	97.04(2.74)
150	100	794.53(821.03)	79.02(13.64)

Further decreases of q did not result in further improvement of the algorithm's performance. In fact, when the value of q was set to 0.5, no simulations resulted in the convergence in 10,000 periods. Regarding the variations in the value of the sensitivity parameter λ, in addition to $\lambda = 0.006$, we conducted sets of simulations with two other values, $\lambda = 0.06$, and $\lambda = 0.6$. None of these simulations converged to equilibrium in 10,000 periods.

We also tested robustness of the results for $|S| = 51$ but, as reported in the text, nothing we did resulted in convergence to equilibrium for $\gamma = 50, 100$, and 150.

5.2. Robustness of Experience Weighted Attraction Learning simulations

In table 8, we report the results of simulations for the values of δ equal to 0.3, 0.5, and 0.7. (Note our baseline simulations were conducted with δ equal to 0.9.) We observed that really high values of δ are required in order for convergence to Nash equilibrium to take place. Lower values of δ resulted in a substantial decrease in the number of simulations that converged to equilibrium. We report the results for $\gamma = 100$.

Table 8
Convergence Times and Stability of Equilibria
EWA with $|S| = 11$, $\gamma = 100$, different values of δ

δ	simulations	$\bar{T}^c_\gamma$ $(\sigma_{T^c_\gamma})$	E^γ_s $(\sigma_{E^\gamma_s})$
0.3	2/100	27.50(20.51)	99.95(0.21)
0.5	7100	6.71 (1.80)	100 (0.00)
0.7	16/100	8.19 (4.83)	99.99(0.05)

A ten-fold decrease in the value of λ (0.035) did not have an impact on the convergence times of simulations with $\delta = 0.9$. However, it did result in a tremendous increase of the number of simulations that converged for lower value of δ. In table 9, we report the results for $\delta = 0.9$, 0.7, and 0.5 combined with 3 values of γ, 50, 100, and 150.

Table 9
Convergence Times and Stability of Equilibria
EWA with $|S| = 11$, $\delta = 0.9$, $N_0 = 10$
Change in the value of $\lambda = 0.035$

$\delta = 0.9$

γ	simulations	$\bar{T}^c_\gamma$ $(\sigma_{T^c_\gamma})$	E^γ_s $(\sigma_{E^\gamma_s})$
50	100	72.64(6.78)	96.86(0.83)
100	100	47.74(12.71)	98.37(0.59)
150	100	64.62(61.86)	99.10(0.47)

$\delta = 0.7$

γ	simulations	$\bar{T}^c_\gamma$ $(\sigma_{T^c_\gamma})$	E^γ_s $(\sigma_{E^\gamma_s})$
50	96/100	66.35(24.92)	99.31(0.42)
100	94/100	38.50(41.15))	99.47(0.35)
150	76/100	50.99(38.33)	99.62(0.36)

$\delta = 0.5$

γ	simulations	$\bar{T}^c_\gamma$ $(\sigma_{T^c_\gamma})$	E^γ_s $(\sigma_{E^\gamma_s})$
50	32/100	62.16(21.74)	99.72(0.26)
100	43/100	44.95(13.38)	99.73(0.29)
150	37/100	40.81(13.38)	99.98(0.05)

Finally, a change in the value of N_0, we tried a ten times lower values of $N_0 = 1$ did not seem to affect the algorithm's behavior. (This is consistent with similar results reported in the literature regarding the robustness of the model to changes in the value of N_0.) The results of our simulations are reported in table 10.

Table 10
Convergence Times and Stability of Equilibria
EWA with $|S| = 11$, $\delta = .9$, $\lambda = 0.35$
Change in the value of $N_0 = 1$

γ	simulations	$\bar{T}^c_\gamma$ $(\sigma_{T^c_\gamma})$	E^γ_s $(\sigma_{E^\gamma_s})$
50	87/100	25.97(31.77	99.95(0.10)
100	100	38.50(41.15))	99.98(0.06)
150	89/100	100.69(133.65)	99.98(0.05)

In case of EWA and $|S| = 51$, we observed no convergence in simulations conducted with lower values of δ (0.3, .4, 0.5, .6, .7, and .8) for $\gamma = 50$, 100 and 150. This confirms the result, obtained in the simulations with $|S| = 11$, that a relatively high value of δ (much higher than what has been reported in the literature) is required for the EWA convergence in the GL environment. This implies that a very high weight has to be placed on hypothetical payoffs if the convergence to equilibrium is to be achieved.

A closer look at the behavior exhibited in the simulations with lower values of δ reveals that in all of the simulations, agents do settle on particular messages whose probability of being selected goes to 1. However, for $\delta = 0.4$ and 0.5, none of the agents settle on equilibrium messages. This also results in lower efficiency observed in these simulations. However, for higher values of δ equal to 0.6 and 0.7 there are simulations where 1 or 2 out of 5 agents settle on equilibrium messages (and this results in higher efficiency). Finally, for $\delta = 0.8$ we observe simulations where 3 out of 5 agents that settle to equilibrium messages.

In table 11 we report the actual values of messages for selected simulations for δ = 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8. The table also reports the value of efficiency at the end of all of these simulations. (The values were taken from the last period, r=10,000, of each simulation). The data show that the simulations do converge to the relatively high values of efficiency even when none of the agents plays equilibrium messages, the values are (at least for the selected set of simulations) above 97%. Efficiency increases to 99% when 2 out of 5 agents play equilibrium messages, and reaches (almost) 100% when 3 out of 5 agents converge to equilibrium messages.

Table 11
Messages, payoffs, and efficiency for different levels of δ
EWA with $|S| = 51$

δ	m^1 (U^1)	m^2 (U_2)	$m^3(U^3)$	m^4 (U^4)	m_5 (U^5)
0.3	1.2 (230.09)	1.6 (192.75)	-0.4 (125.25)	0.8 (226.69)	0.6 (229.97)
0.4	1.4 (237.96)	2.0 (166.08)	- 0.4 (128.72)	0.8 (250.96)	0.4 (237.84)
0.5	1.4 (209.16)	1.8 (209.68)	1.2 (207.12)	1.2 (206.16)	0.6 (172.64)
0.6	1.6 (190.83)	1.0 (231.81)	1.2 (202.85)	1.2 (203.49)	0.8 192.51)
0.7	1.2 (202.83)	0.8 (228.61)	0.8 (200.05)	1.0 (201.09)	1.0 (201.573)
0.8	1.0 (206.33)	0.8 (228.00)	1.0 (201.33)	1.2 (198.00)	1.0 (201.33)

Table 9, cont.d

δ	$\sum U$	efficiency
0.3	1004.76	0.971
0.4	1021.56	0.987
0.5	1004.76	0.971
0.6	1021.56	0.987
0.7	1034.16	0.999
0.8	1035	1.000

In table 12 we report the equilibrium messages and payoffs for individual agents and the sum of equilibrium payoffs for the purpose of comparing it with the data in table 11. This comparison reveals that in thee simulations there is always 1 or 2 agents with payoffs higher than their equilibrium payoffs, while the est of the agents end up with payoffs lower than their equilibrium ones.

Table 12
Equilibrium messages and payoffs in case of $\gamma = 100$

m^1 (U^1)	m^2 (U_2)	m^3(U^3)	m^4 (U^4)	m_5 (U^5)	$\sum U$
1(205)	1 (230)	1 (200)	1 (200)	1 (200)	1035

A ten-fold decrease in the value of λ from 0.35 to 0.035 resulted in a slow down of the time to convergence. We conducted 100 simulations for $\gamma = 50$, 100 and 150. A ten fold decrease in the value of the response sensitivity parameter λ resulted in a substantial increase in the number of simulations that converged to equilibrium. As mentioned above, this low value of λ slowed down convergence of EWA with $|S| = 11$ resulting in convergence times that were longer than what the experimental evidence suggests. However, in the case of EWA with $|S| = 51$, this low value helps the algorithm achieve convergence in a much larger number of simulations. However, convergence times are still much longer than those observed in the experiments. We report the average convergence times and measures of stability in table 13.

Table 13
Convergence Times and Stability of Equilibria
EWA with $|S| = 51$ and $\lambda = 0.035$

γ	simulations	$\bar{T}^c_\gamma$ ($\sigma_{T^c_\gamma}$)	E^γ_s ($\sigma_{E^\gamma_s}$)
50	0		
100	68/100	1228.69 (1009.61)	89.49 (1.63)
150	41/100	967.59 (1157.71)	91.54 (1.42)

Combinations of a decreased value of λ and values of δ less than 0.9 did not result in an increase in the number of simulations that converged to equilibrium.

We also examined the sensitivity of the algorithm to the increases in the value of λ. We simulated the model for the values of λ equal to 2, 5, 10, 15, ad 18 (keeping the other parameter values equal to our baseline setting) and for $\gamma = 100$. None of these simulations converged to the Nash equilibrium.

Decreasing the value of $N(0)$ from the baseline value of 10 to 1 did not result in a significant change in terms of the ability of the algorithm to converge. Only one simulation converged for $\gamma = 50$, 3 converged for $\gamma = 100$, and 1 converged for $\gamma = 150$. Other studies also suggest that differences in this value do not have significant effects on the EWA's behavior. We report the results in table 14.

Table 14
Convergence Times and Stability of Equilibria
EWA with $|S| = 51$ and $\lambda = 0.35$ and $N_0 = 1$

γ	simulations	$\bar{T}^c_\gamma$ $(\sigma_{T^c_\gamma})$	E^γ_s $(\sigma_{E^\gamma_s})$
50	1/100	38	100 (0)
100	3/100	32.67 (2.08)	99.93 (0.12)
150	1/100	31	100 (0.00)

Finally, we varied ϕ above and below the basic value of 0.9. The results are displayed in Table 15. As one can see, there is relatively difference between the values of 0.7 and 0.9. But both the increase to 0.99 and decreases to 0.5 and below slow things down.

Table 15
Convergence Times and Stability of Equilibria
EWA with $\gamma = 100$,$|S| = 51, \lambda = 0.35$ and $N_0 = 10$

ϕ	simulations	$\bar{T}^c_\gamma$ $(\sigma_{T^c_\gamma})$
.99	100	25.09(28.68)
100	100	32.67 (2.08)
.9	100	12.9(6.98)
.7	100	12.37(4.46)
.5	0/100	DNC
.3	0/100	DNC

DNC =Did Not Converge in 100 trials

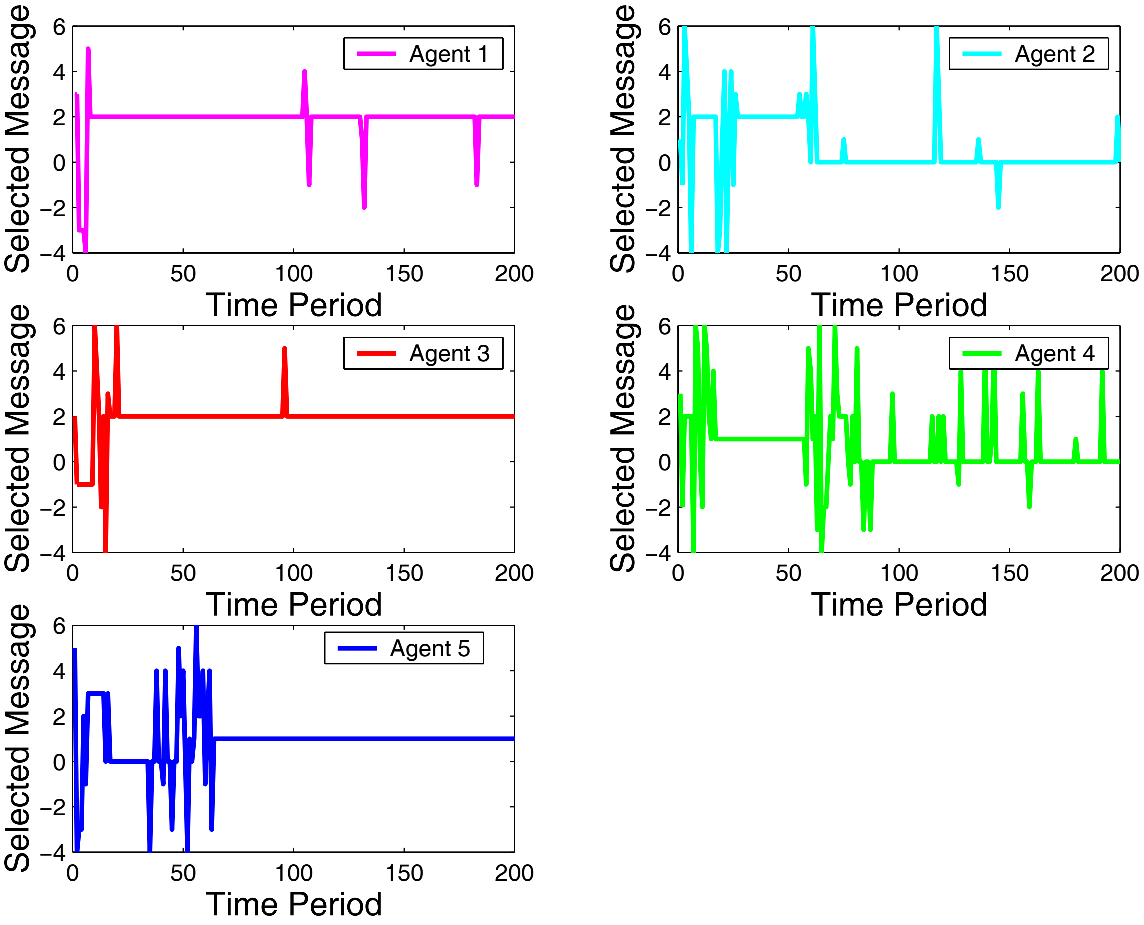

Figure 1. Behavior of selected messages for RL algorithm and S=11

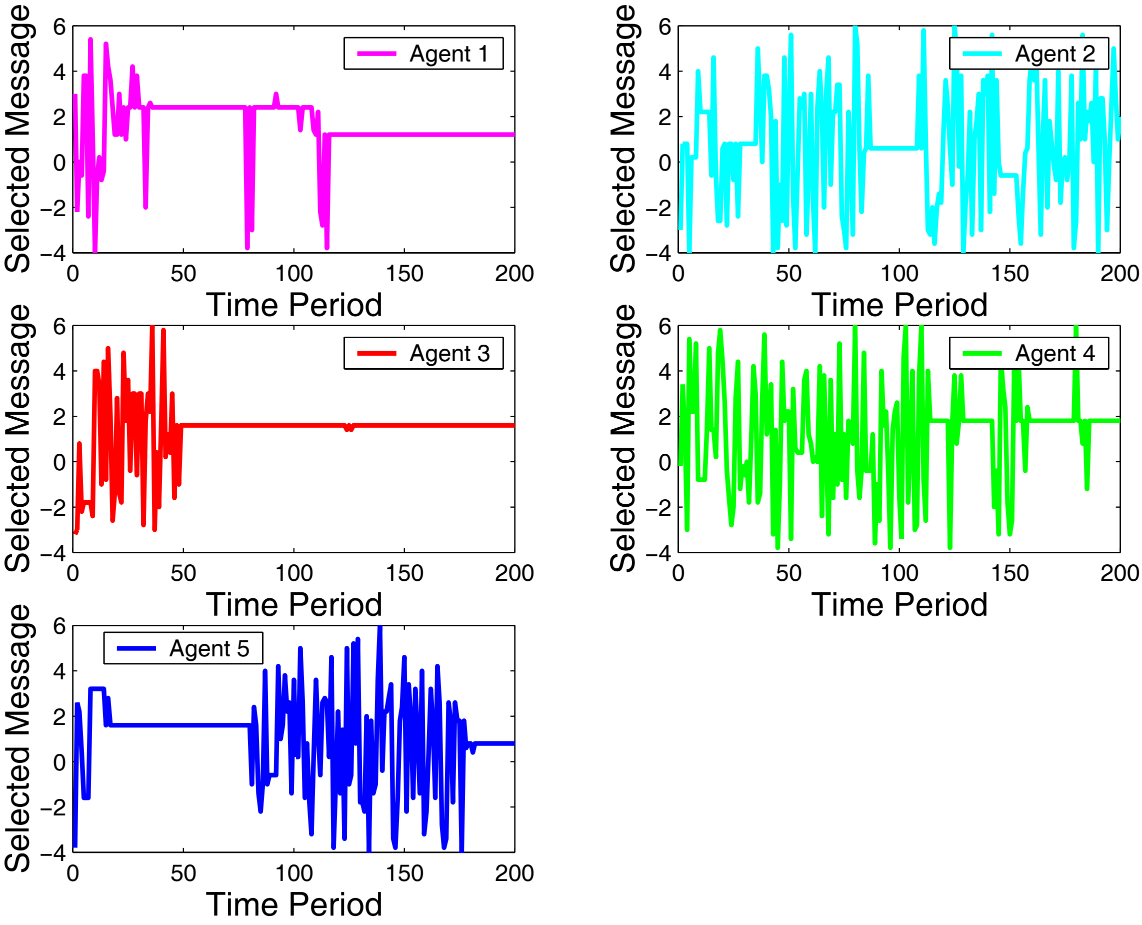

Figure 2. Behavior of selected messages for RL, $|S| = 51$

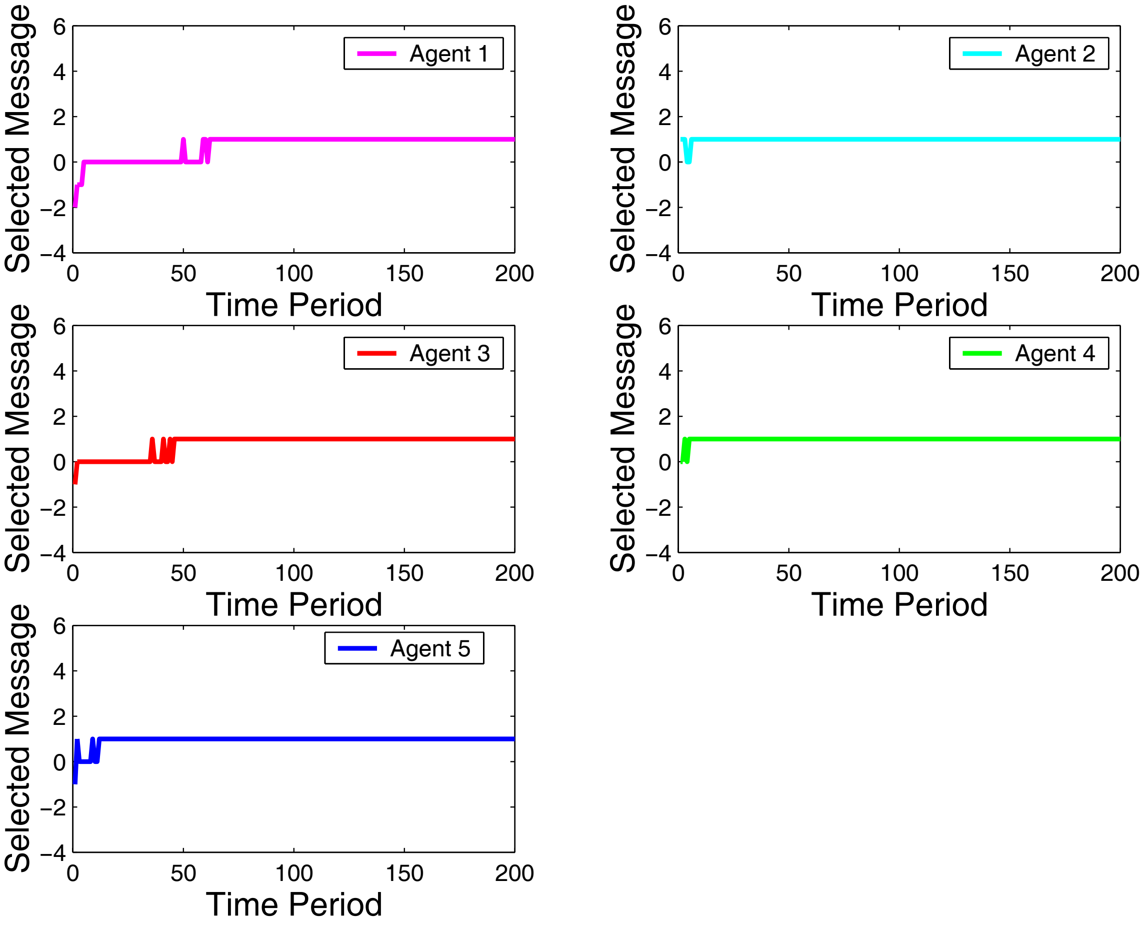

Figure 3. Behavior of selected messages for EWA, $|S| = 11$

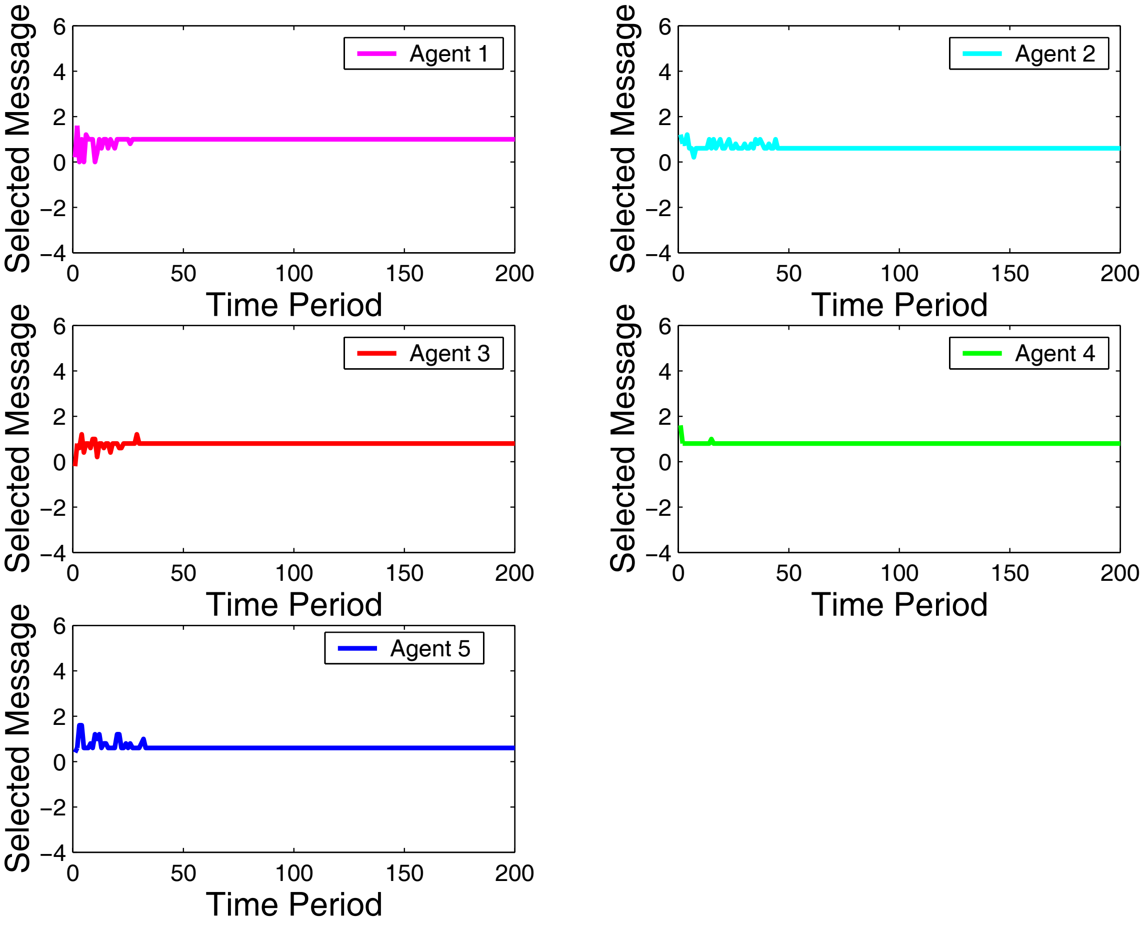

Figure 4. Behavior of selected messages for EWA, $|S| = 51$

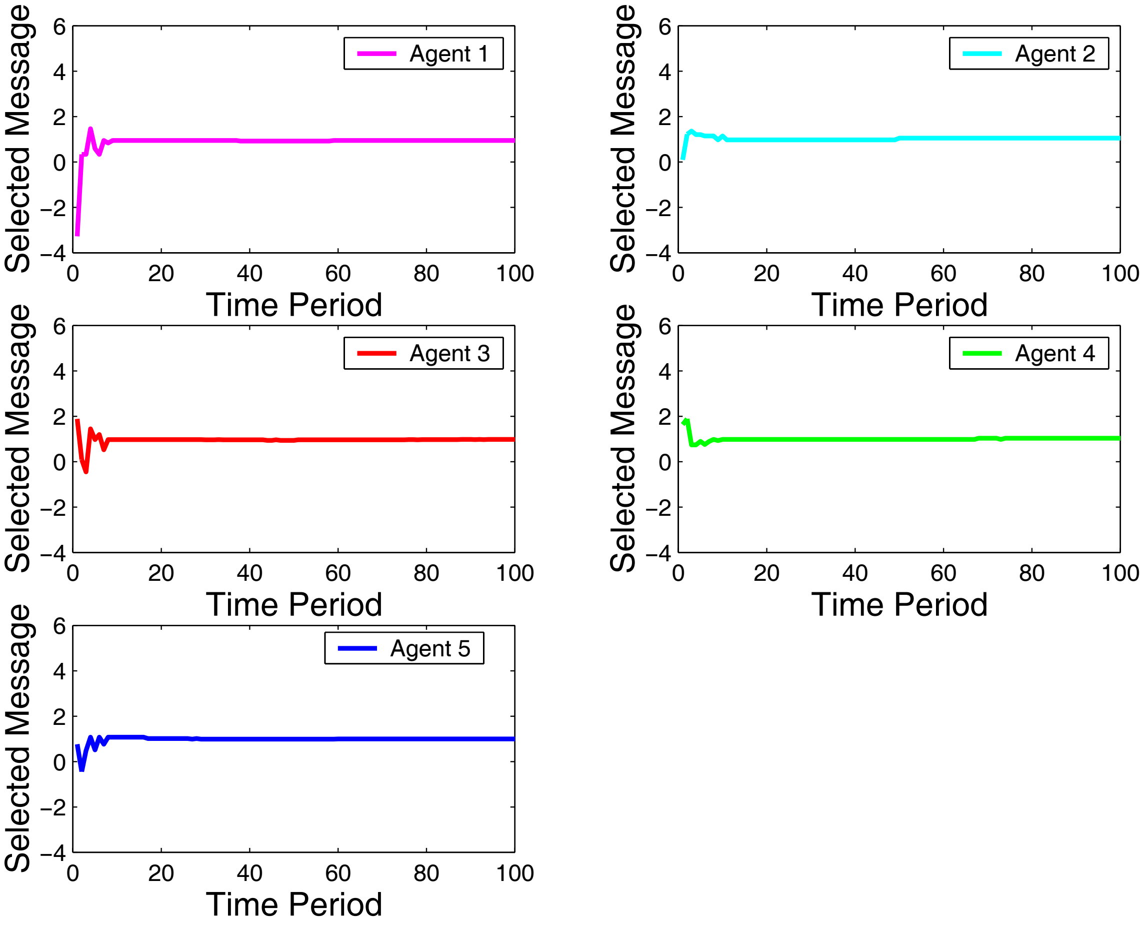

Figure 5. Behavior of selected messages for IEL
J=100, $p_{ex} = 0.033$, normal distribution

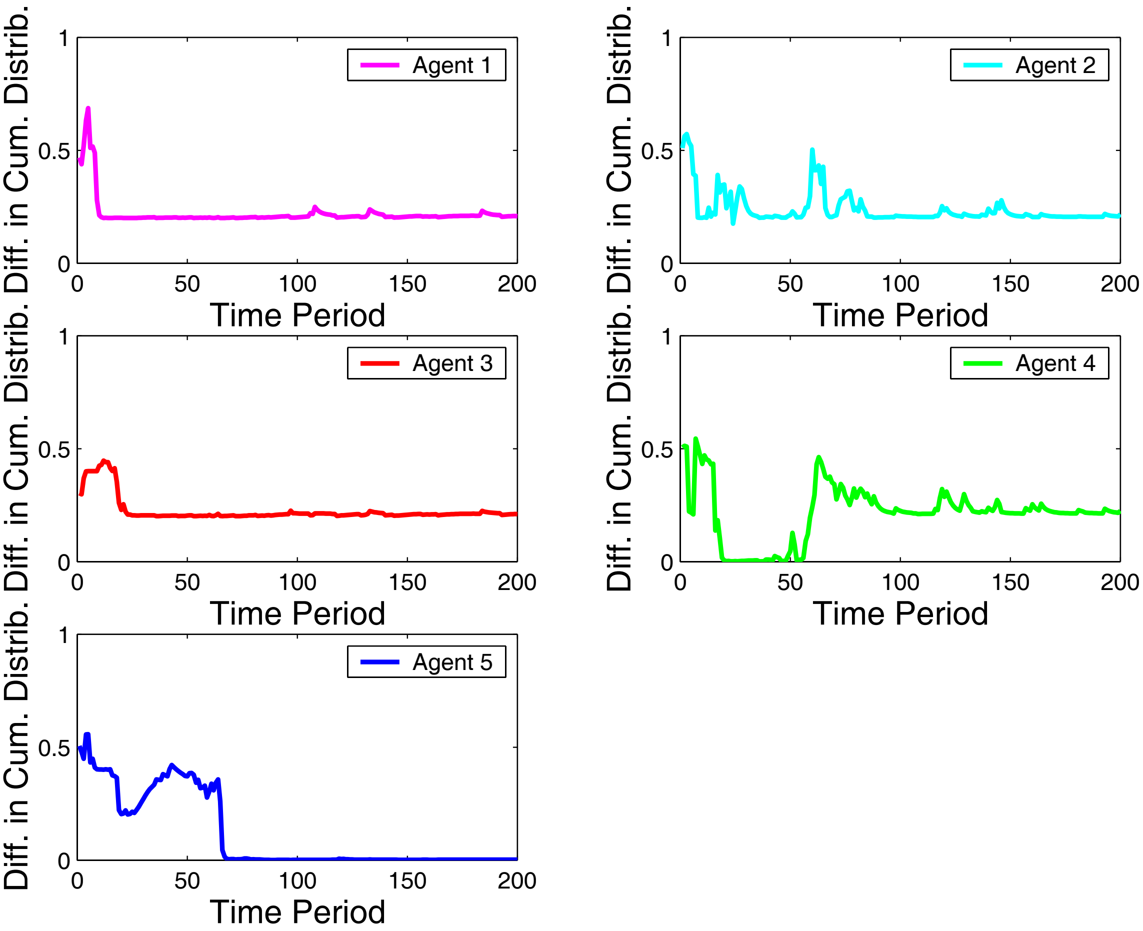

Figure 6. Difference in cumulative probability distributions for RL, $|S| = 11$

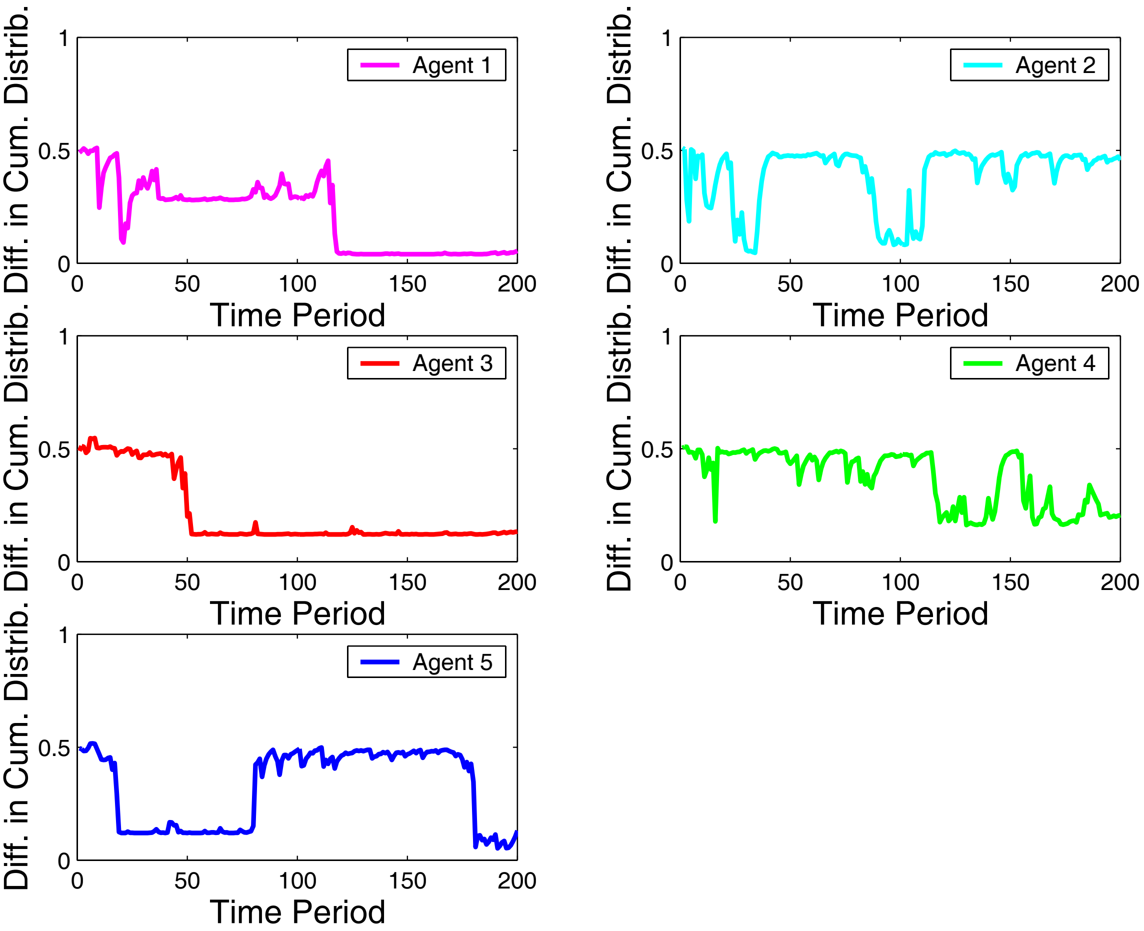

Figure 7. Difference in cumulative probability distributions for RL algorithm, $|S| = 51$

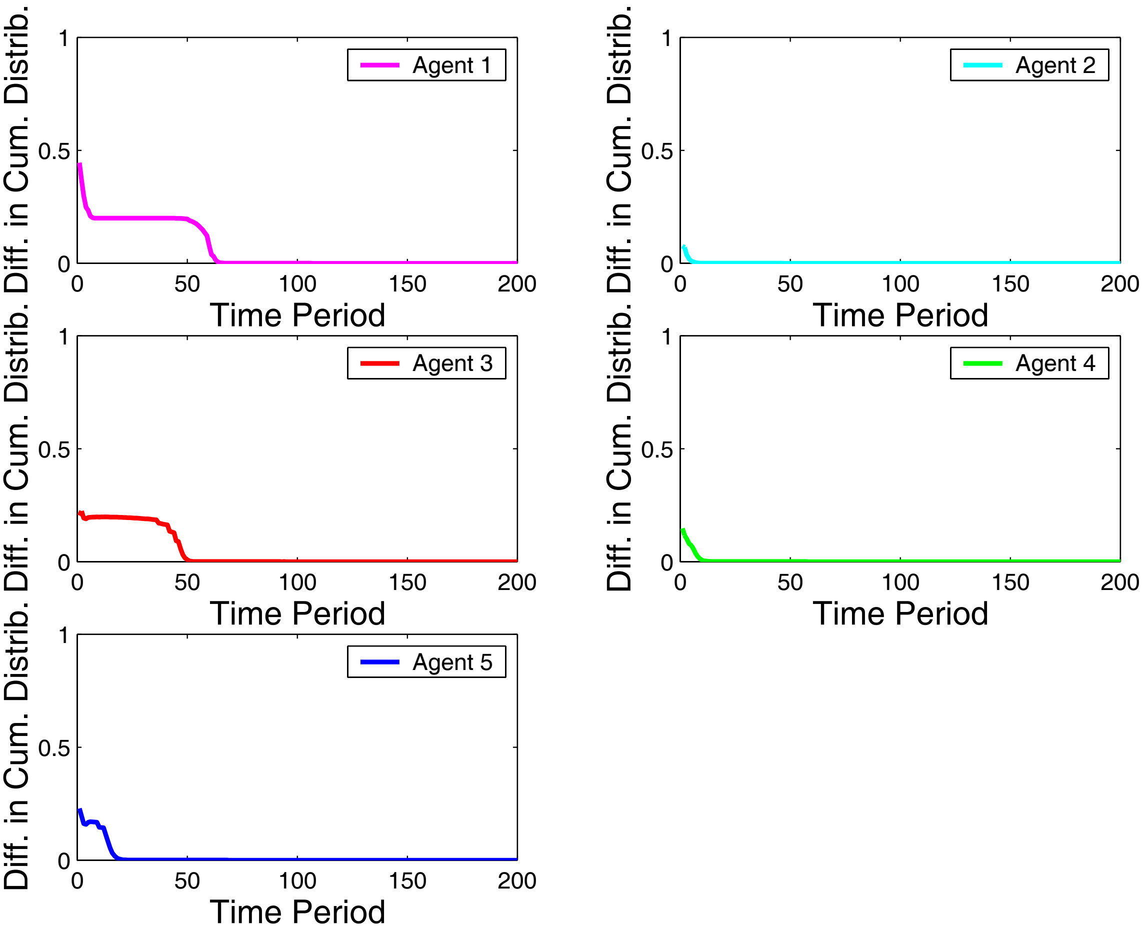

Figure 8. Difference in cumulative probability distributions for EWA, $|S| = 11$

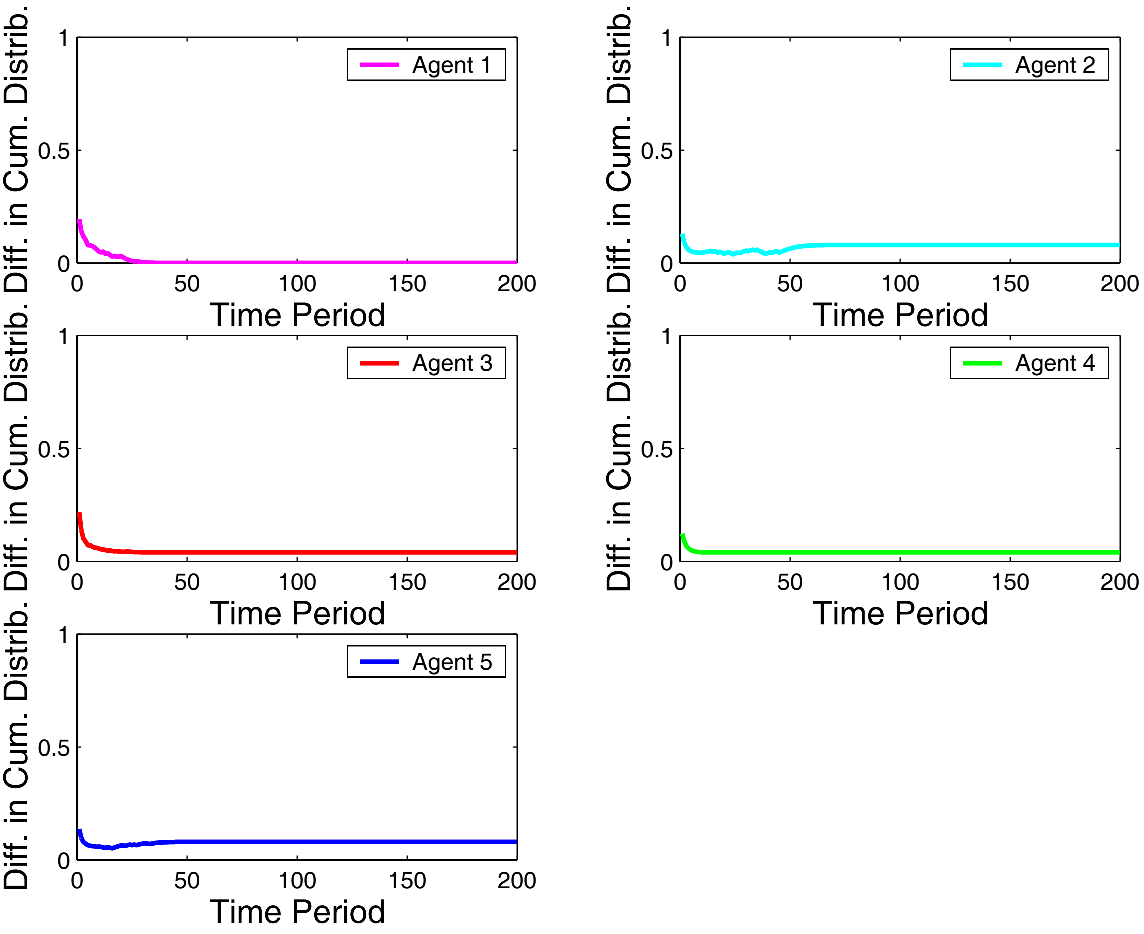

Figure 9. Difference in cumulative probability distributions for EWA , $|S| = 51$

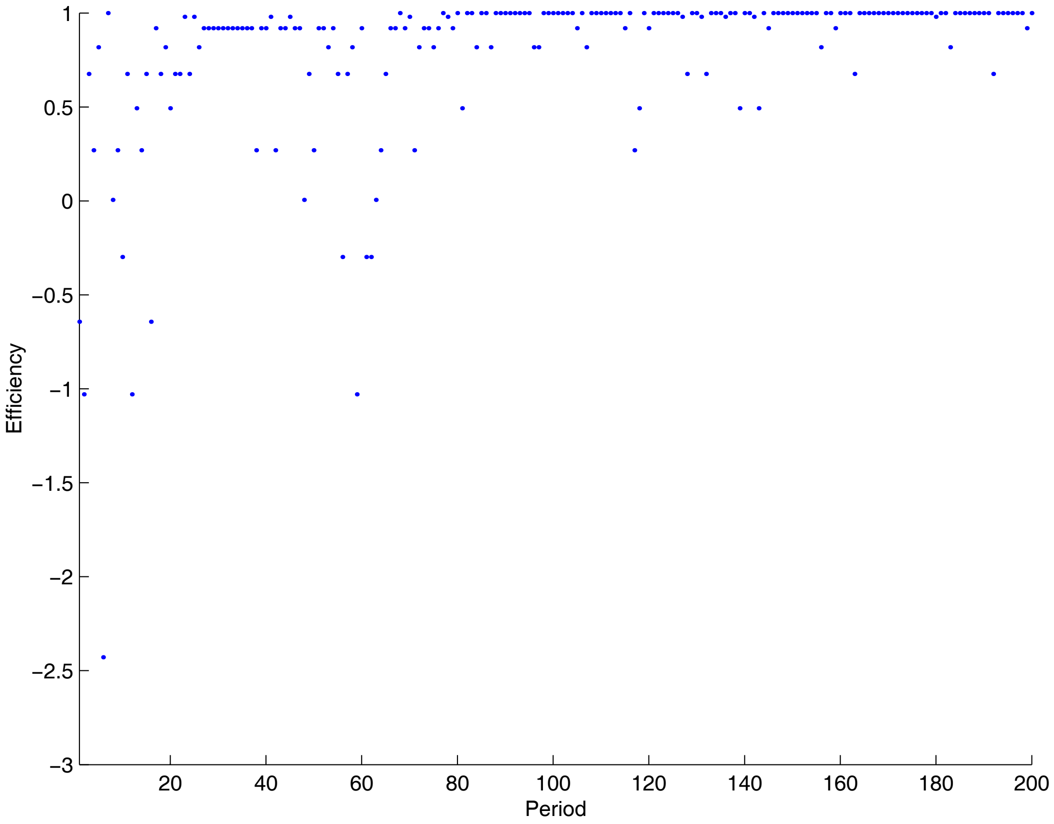

Figure 10. Efficiency for RL algorithm, $|S| = 11$

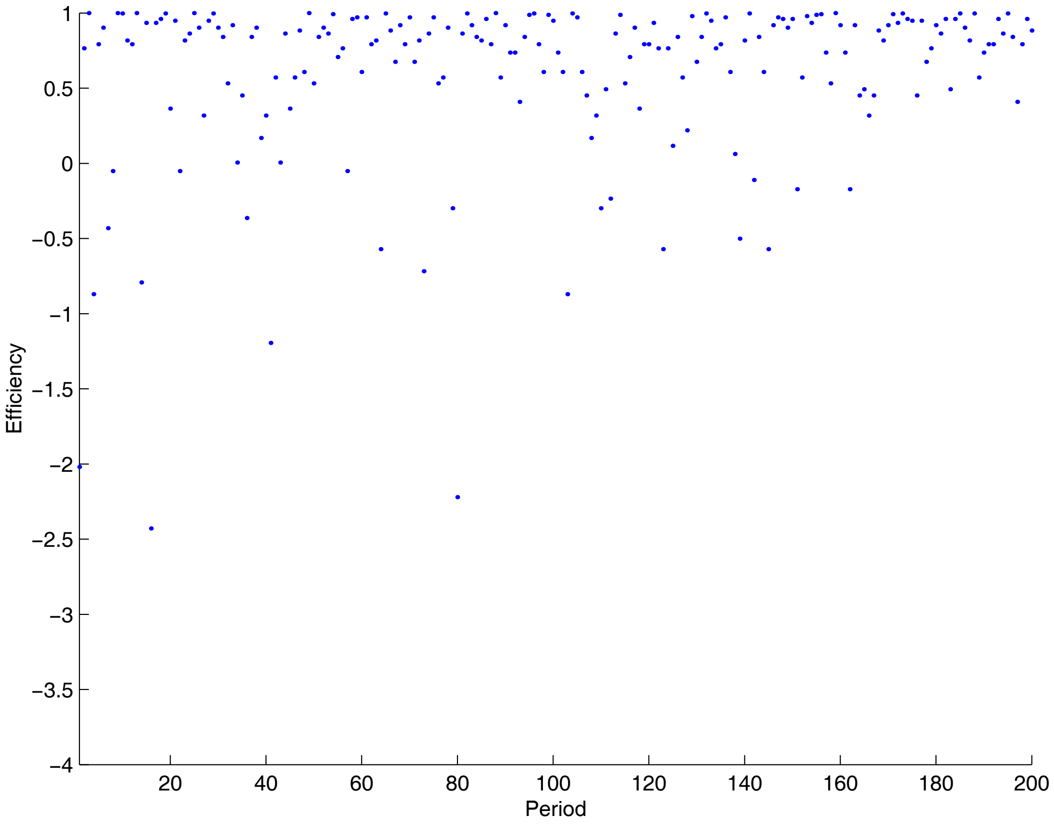

Figure 11. Efficiency for RL algorithm, $|S| = 51$

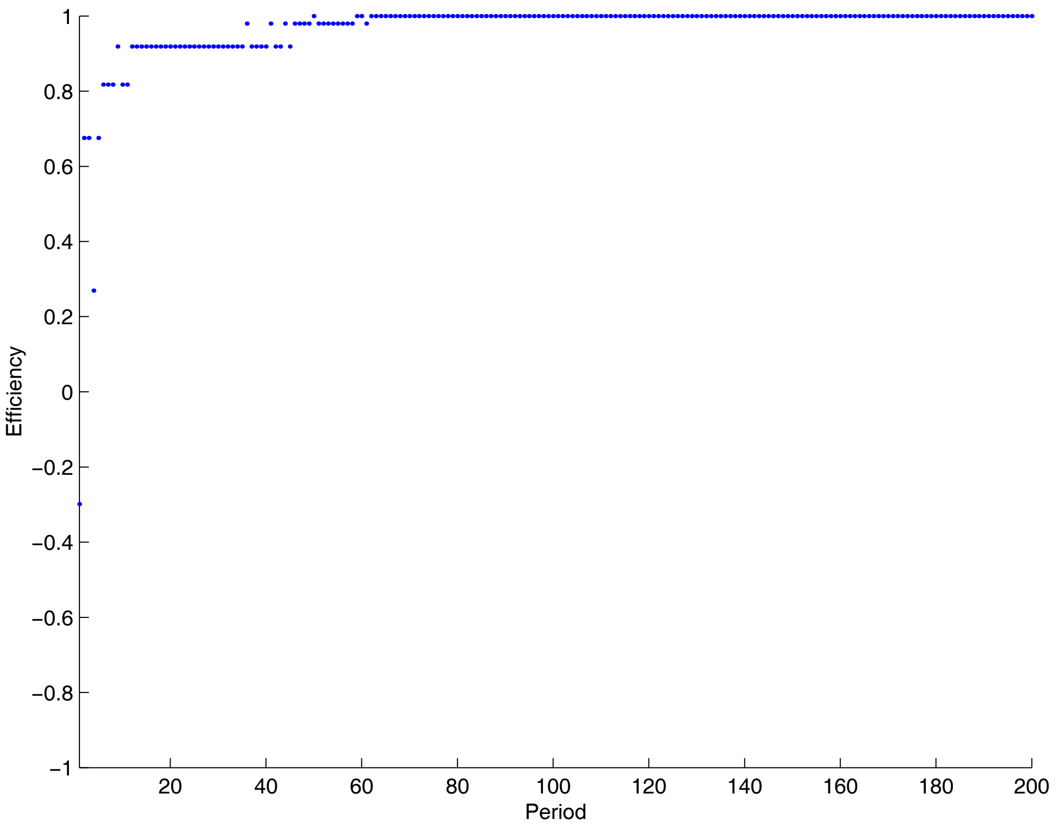

Figure 12. Efficiency for EWA, $|S| = 11$

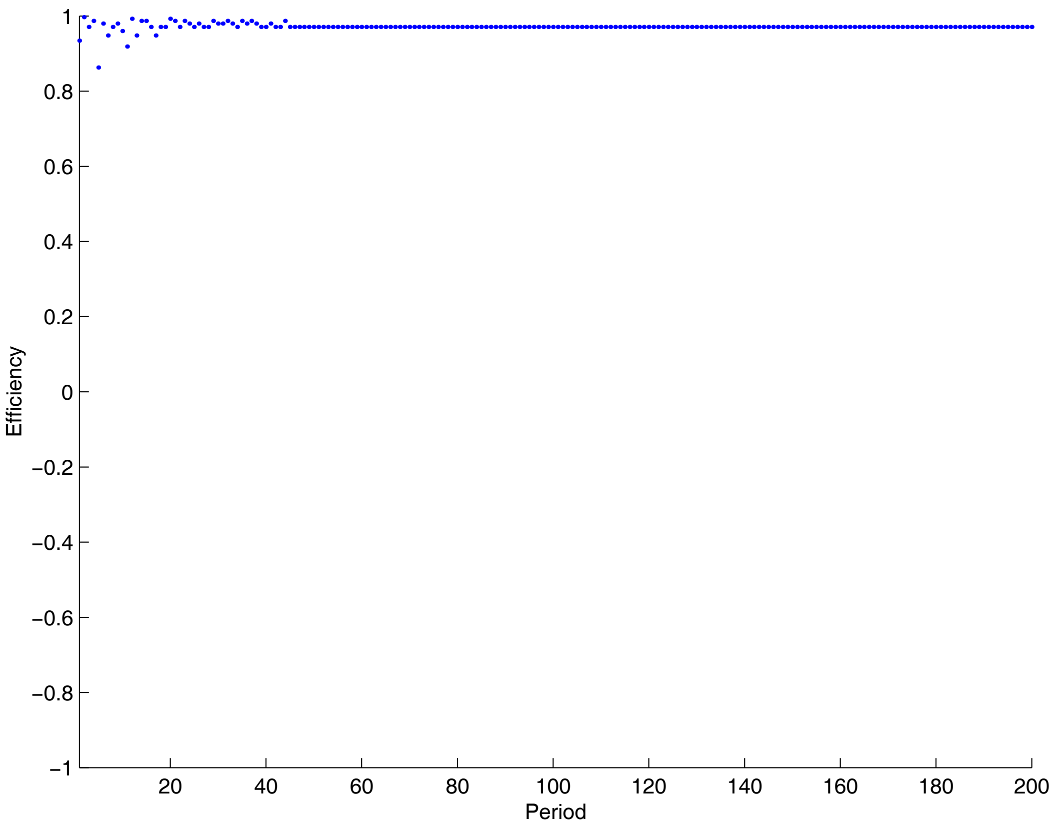

Figure 13. Efficiency for EWA, $|S| = 51$